Individual Consequences and Public Challenges of Social Change

–

Six Empirical Essays in Health Economics and Economics of Education

Der Fakultät für Wirtschaftswissenschaften der
Universität Paderborn

zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
– Doctor rerum politicarum –
vorgelegte Dissertation

von

Matthias Westphal, Master of Science
geboren am 27.02.1987 in Münster

Paderborn, 2017

Preface

In the first place, I want to express my gratitude to my supervisor Hendrik Schmitz. From the point when I became part of his team and over the years, I benefited substantially from his input and from his view on applied econometric research. In the beginning, he provided me with the necessary freedom to incorporate my own ideas on a rather elaborated research question, while guiding me to trim the resulting paper towards scientific success. This process helped me to evaluate the general potential of research questions especially with respect to publishing in scientific journals. In later stages of my scientific training, he gave me confidence in believing in the value of my ideas and encouraged me to frame them such that one idea eventually got funding from a nonprofit foundation.

Furthermore, I also gained a lot from Martin Karlsson. He created a competitive environment with the research seminars and the annual conference, where it was possible to compare to the top and to get advice from many high-quality researchers from versatile areas of applied microeconometrics.

Daniel Kamhöfer, my coauthor and primary office colleague, likewise belonged to Hendrik Schmitz' team. Beside being a great coauthor who complemented my skills, his work ethic and passion for writing research papers definitely rubbed off on me. I also want to acknowledge the impact of Martin Fischer who fueled my interest in formal and more complex econometrics and methods of program evaluation alike. He additionally co-initiated the Essen empirical reading group that sparked the idea of studying the marginal treatment effect in detail and, thereby, laid the foundations for the research questions that are discussed in two chapters of the thesis at hand. Moreover, I would like to point out the contribution of my remaining co-authors, Claudia Andreella and Therese Nilsson. I also thank my second supervisor Wendelin Schnedler for not hesitating to step in on rather short notice as my second supervisor. Likewise, I appreciate the support of the RWI and its researchers, foremost that of Ansgar Wübker.

I wrote this thesis while I was student at the Ruhr Graduate School in Economics. Here, I benefited considerably from the research atmosphere, the infrastructure, and the people engaged in and being part of the RGS, in particular my colleagues and friends from the 10th cohort.

Furthermore, I am thankful for the high-quality input from the outside my direct research network. This not only comprises all those who read, revised, and/or commented my papers. It also and particularly includes comments from many referees, as the current dissertation benefited a lot from peer review. Additionally, I want to acknowledge monetary input, namely the fundings that made the realization of all the

studies included in this dissertation possible: the Fritz-Thyssen Stiftung that funded the work on three chapters in two different projects and the Deutsche Forschungsgemeinschaft that funded the work on two additional chapters.

Last, but certainly not least, I am deeply indebted to my girlfriend Lisa König who, over the years, proved much patience and stamina. She took the load off me and supported me whenever possible.

Contents

I Introduction **1**

II Consequences of the demographic transition **17**

1 The long shadows of past insults – intergenerational transmission of health over 130 years **19**

1.1 Introduction . . . 19

1.2 Literature review . . . 23

1.3 Data . . . 25

1.3.1 Individual level data . . . 25

1.3.2 Maternal disease environment in utero . . . 29

1.3.3 Economic environment . . . 30

1.4 Empirical strategy . . . 30

1.5 Results . . . 32

1.5.1 Explorative graphical analysis . . . 32

1.5.2 Second generation mortality . . . 33

1.5.3 Second generation SES outcomes . . . 40

1.5.4 Tracing the health insult . . . 47

1.6 Conclusion . . . 52

1.7 Appendix . . . 54

2 Short- and medium-term effects of informal care provision on female caregivers' health **59**

2.1 Introduction . . . 59

2.2 Institutional background . . . 61

2.3 Empirical strategy . . . 62

2.4 Data 66
2.5 Results 70
2.5.1 Matching quality 70
2.5.2 Estimation results 72
2.6 Sensitivity analysis 75
2.7 Conclusion 77
2.8 Appendix 78

3 Informal care and long-term labor market outcomes 91
3.1 Introduction 91
3.2 Institutional background 93
3.3 Data 94
3.3.1 Sample selection 94
3.3.2 Informal care paths 96
3.3.3 Outcome variables 98
3.4 Baseline analysis 99
3.4.1 Empirical strategy I – A static design 99
3.4.2 Control variables 101
3.4.3 Potential failure of the CIA 103
3.4.4 Estimation results – Static model 103
3.5 Sensitivity analysis 106
3.5.1 General framework 107
3.5.2 Choice of selection and outcome effects 108
3.5.3 Results 110
3.6 A dynamic design 111
3.6.1 Empirical strategy II 111
3.6.2 Estimation results – Dynamic model 114
3.7 Alternative specifications of the treatment variable 116
3.8 Conclusion 117
3.9 Appendix 119

III Consequences of the educational expansion 139

4 Heterogeneity in marginal non-monetary returns to higher education 141

4.1 Introduction . . . 141
4.2 Institutional background and exogenous variation . . . 143
4.2.1 The German higher educational system . . . 143
4.2.2 Exogenous variation in college education over time . . . 144
4.3 Empirical strategy . . . 148
4.4 Data . . . 152
4.4.1 Sample selection and college education . . . 152
4.4.2 Dependent variables . . . 152
4.4.3 Control variables . . . 155
4.4.4 Instrument . . . 156
4.5 Results . . . 157
4.5.1 OLS . . . 157
4.5.2 Marginal treatment effects . . . 159
4.5.3 Treatment parameters . . . 162
4.6 Potential mechanisms for health and cognitive abilities . . . 165
4.7 Conclusion . . . 168
4.8 Appendix . . . 170

5 Fertility effects of college education: evidence from the German educational expansion 185

5.1 Introduction . . . 185
5.2 Trends in fertility and education in Germany . . . 188
5.3 The college expansion . . . 190
5.3.1 Background and developments . . . 190
5.3.2 Determinants of the college expansion . . . 191
5.4 Data and empirical strategy . . . 194
5.4.1 Survey data and important variables . . . 194
5.4.2 Empirical strategy . . . 198
5.5 Baseline results . . . 200

5.5.1 The effect of the college expansion on educational participation 200
5.5.2 The effect of college education on fertility 203
5.6 Heterogeneity and potential mechanisms 205
5.6.1 Effect heterogeneity along age 205
5.6.2 Opportunities and revealed preferences for career and family life 210
5.7 Conclusion 212
5.8 Appendix 214

6 More teachers, smarter students? Potential side effects of the German educational expansion 219
6.1 Introduction 219
6.2 The market for teachers in Germany 222
6.3 Empirical strategy and theoretical mechanism 224
6.3.1 Empirical strategy 224
6.3.2 Theoretical mechanism 227
6.4 Data 231
6.4.1 Sample selection and student-teacher linking 231
6.4.2 Test score data 232
6.4.3 Descriptive statistics 233
6.5 Results 234
6.5.1 Effects on students' learning outcomes 234
6.5.2 Assessing the validity of the estimates 238
6.5.3 Detecting teacher selection in the characteristics of teachers . 240
6.6 Conclusion 243
6.7 Appendix 244

IV Concluding remarks 255

Bibliography 260

List of Tables 279

List of Figures 281

Part I

Introduction

Introduction and motivation

Over the last decades, many fundamental aspects of our lives changed substantially. Compared to the generation of our grand- and great-grandparents, for instance, we can today live both longer and in better health, are much higher educated, and typically, we live together in considerably smaller families. Many of these changes are interrelated. And all have transformed individual decisions over the life course and have thus changed how the society is (or needs to be) organized. Especially two developments of the past 60 years describe and comprise many of these transformations: the demographic transition – characterized by the rapid aging of many industrialized societies (Harper, 2014) – and the educational expansion – describing the upsurge in educational attainment rates (see Schofer and Meyer, 2005 and Goldin and Katz, 2009).

The origins of the demographic change and the educational expansion are multifaceted – and yet intertwined. The former, the population aging, is caused by two distinct developments. On the one hand, longevity increased due to medical innovations and knowledge about health consequences of individual behavior (Cutler et al., 2006). On the other hand, innovations in contraceptive techniques and, maybe more important, a rise in female educational attainment and labor force attachment (Goldin, 2006) caused fertility rates to continue to decrease significantly among industrialized societies (Sobotka, 2004). The other substantial development of the past six decades – the educational expansion – was driven by a century of skill-biased technological change that increased the monetary returns to education (Acemoglu, 2002; Autor, 2014; Goldin and Katz, 2009). Beside this monetary aspect, also a non-monetary virtue contributed to upsurging educational enrollment rates: education facilitates social and political participation – a notion that gained in importance and shaped the public opinion from the early 1960s onward (Dahrendorf, 1965).

Moreover, and this unites the educational expansion with the demographic transition, there are spill-over effects from other social developments. Four reasons make the educational expansion to be closely interrelated with the demographic change. First, an increased life expectancy raises the individuals' incentive of educational investments – a mechanism, which has been formalized and introduced to the economic discussion by Ben-Porath (1967) and empirically assessed by Bleakley (2017), among others. Second, to mitigate adverse consequences of the demographic change, education has been proposed as an important means to reduce the fiscal and economic strain of population aging (see European Union, 2006 and Börsch-Supan, 2003). Third, education can reduce the consequences of population aging for individuals by affecting health or health behavior (see Grossman, 2006 or Oreopoulos and Salvanes,

2011). As one mechanism, for instance, education can postpone people from suffering from the most frequent cause of why people become care dependent[1] – dementia – by (potentially) increasing cognitive skills or the cognitive reserve (Brayne et al., 2010).[2] This is propagated by the cognitive reserve hypothesis (see Alzheimer's Association, 2017 or Stern, 2012). Lastly, and in contrast to the preceding arguments, the educational expansion can also magnify the aging of the society. Increased female labor participation and educational attainment increased the opportunity costs of childbearing (Adda et al., 2017), which gives rise to the so-called "baby-gap", the female fertility gradient in education (Raute, 2017). I focus in this thesis on this intimate interrelation between aging and education by dedicating two chapters to this nexus.

The general goal of this dissertation is the assessment of individual and societal consequences that are caused by the changes specified and characterized above. To begin with, I start by generally analyzing how long it may take for a society to adapt to a specific social change, the epidemiological transition, that was at the root of the demographic transition (Chapter 1). This epidemiological transition was a major cause of increased longevity in the last 150 years and is characterized by a significantly declining mortality rate due to infectious diseases, especially among infants (Cutler et al., 2006). By looking at how societal or health inequalities that are caused by infectious diseases transmit from one generation to the next, I assess whether the rate of adaption to social changes is slower as commonly suspected. According to the underlying hypothesis, the mortality patterns in today's societies may be attributed to the disease environment that prevailed generations ago. Hence, this would have implications for societal inequalities in the very long-run. Public interventions thereby may generate returns over a very long time period, an important aspect that this chapter tries to shed light on.

Then, I turn to some particular direct and immediate consequences of the aging of the population: the increased demand for long-term care (in Chapters 2 and 3). At the heart of the public debate on long-term care is how the considerably expanded demand in future will be met. By increased formal care in nursing homes or by informal care provided, for instance, by close relatives. When it comes to this supply side, most European societies have a preference for informal long-term care (in Germany, two thirds of all care dependents are exclusively cared by relatives). Apart from being preferred by most care recipients, it also imposes less direct costs on the long-term care system. Yet, there may also be indirect fiscal costs. Increased female labor force participation, as being among the means to reduce the fiscal strain of population aging, might increase the burden of care among women who traditionally provide informal long-term care (LTC) services. Insights into the individual effects of individuals who take up this burden and care informally are highly informative for assessing the indirect costs of informal care. Yet, the assessment of the complete indirect costs requires knowledge about the evolution of these effects over time. For example, these

[1] Half of all caregivers in the US provide care to people with dementia (Alzheimer's Association, 2017).

[2] Alzheimer's disease, in turn, is the most common cause of dementia. In the US, 5.5 million individuals in 2016. It's prevalence clearly is age-dependent. Three percent of people age 65-74, 17 percent of people age 75-84, and 32 percent of people age 85 or older have Alzheimer's dementia (Alzheimer's Association, 2017).

costs may be reflected in adverse health effects or reduced labor supply of informal caregivers that might go beyond the caregiving spell. Such indirect costs may mitigate or even magnify the fiscal strain. How exactly these indirect costs affect the fiscal burden is an empirical question, which will be addressed in this thesis.

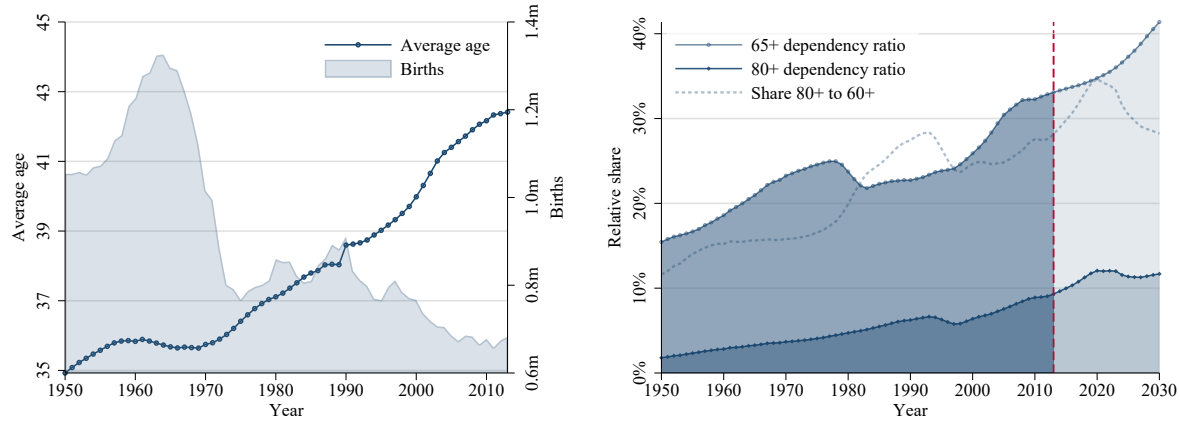
Subsequently, I shift the focus slightly away from direct repercussions of the demographic change to education – the other driver of social changes in the past decades. The effects of education are highly informative also and especially in the context of the demographic change. On the one hand, education is said to act as a preventive measure against the consequences of demographic change. On the other hand, it is also suspicious of being one of its driving forces. To disentangle the effect of education on the demographic change comprehensively, I focus on both the prevention of the effects of demographic aging and, in contrast, on magnifying the population aging. Concerning the former, I evaluate the effects of college education on health and cognitive skills (Chapter 4). Regarding the latter, I comprehensively analyze the effects of college education on fertility decisions of women (Chapter 5).

Lastly, to bring the analysis of these changes full circle, I return to a more general perspective to assess the side-effects of policy-induced changes (Chapter 6). Policy-induced changes, such as the educational expansion, may become necessary because of latent social changes. Therefore, I address the question whether demand shocks in the labor market of teachers during the educational expansion encouraged individuals of lower teaching abilities to become teachers. The identification of a tradeoff between teacher quality and quantity is highly informative since potential effects on students' outcomes are likely to be persistent. It is policy-relevant because these effects are malleable by, for instance, more restrictive teacher hiring policies. Although this provides direct evidence on the overall effects of the educational expansion, the results resort to long-term care, since the effects may be well-transferable into general effects of social changes, such as potential side-effects of a rapid expansion of the nursing homes spots or the recent enlargement of the daycare system for childcare.

Some descriptives

The demographic transition

As declining working age populations generally jeopardize all pillars of social security systems that are organized as pay-as-you-go systems, the demographic transition imposes many challenges in particular to European societies. Figure 1 illustrates prominent descriptives that characterize the demographic change in Germany. The right panel shows the evolution of the mean age. Whereas the average German was aged 35 in 1950, the population aged by more than 7 years to reach 42.4 years in 2013. This has causes as well as consequences. Among the direct causes of the population aging are plummeting fertility rates, which are depicted in the background of this panel. After some baby-booming years at the beginning of the 1960s, the number of births declined almost continuously. The number of annual births nearly halved to date, falling from nearly 1.4 million to slightly above 700 thousand, which gave rise to the suspicion that this was in part driven by the simultaneous upsurge in education, as we shall see later in this section.



(a) The average age by year

(b) The old age dependency ratio

Figure 1: Descriptives of the demographic change in Germany (including GDR)

Notes: All these statistics refer to the German population with the borders of 1990. The small discontinuous jump in 1990 may be attributed to adjusted methods of measurement. The 65 plus old age dependency ratio is defined to be the fraction of of individuals over the age of 64 compared to individuals aged 15 to 64. Accordingly, the 80 plus dependency ratio is the fraction of of individuals over the age of 79 also compared to individuals aged 15 to 64. After 2013, the depicted values are prediction under the assumptions of a constant number of births, and constant migration and mortality. Source: own calculations using official statistics provided in Franzmann (2015).

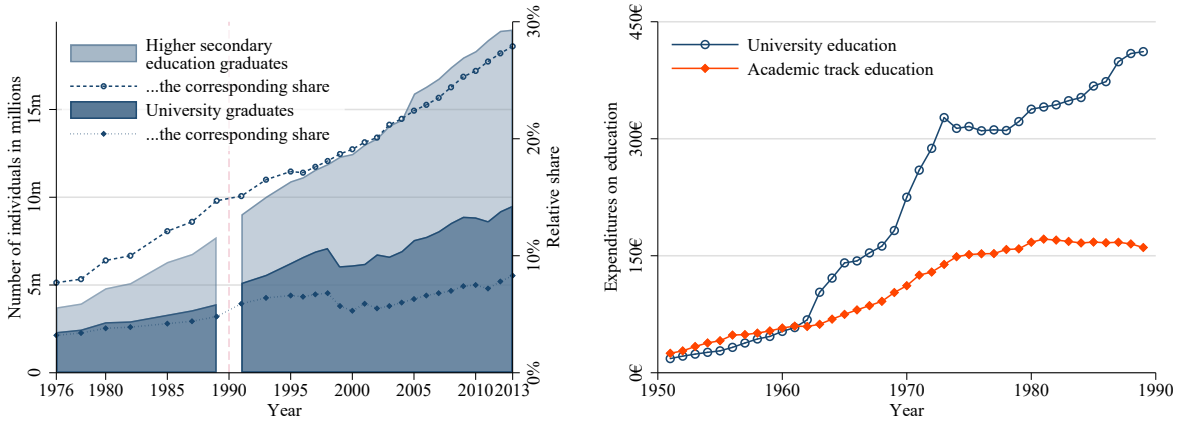
One of the consequences of the demographic change is clarified in the second panel. Here, two versions of the old age dependency ratio are presented. The first version (65+) is the the fraction of people above the age of 64 to people in working age (15 to 64), which makes the demographic change particularly apparent. From 1950 to 2013, this number more than doubled, climbing from 15.4 to 33.1 percent. Until 2030, this number is predicted to rise further to reach 41.3 percent under the assumption of constant births (which does not affect the prediction in the first 15 years), mortality and migration. As these dependency rates approximately show the number of benefit recipients to contributors in a pay-as-you-go social security system such as Germany, these statistics also demonstrate the fiscal consequences of the demographic transition. The upsurging fiscal expanses have implications for the pension system in the first place and the labor market in general (Harper, 2014) . However, the direct consequences are mitigated by an overall increasing labor force, due to many women who entered the labor market.

The other direct implications of aging concern health and long-term care. While it is less clear to which extent population aging is responsible for the rise in general health care expenditures (see, for instance, the arguments in Breyer et al., 2015 versus Zweifel et al., 1999), the most direct consequence of aging societies is an increased demand for LTC services (see, e.g, Karlsson and Klohn, 2014 and Werblow et al., 2007). As LTC expenditures and dependency ratios exponentially increase in age, it is no surprise that age-related LTC expenditures as measured relative to the GDP are expected to have doubled (from the current level of 1.8 percent of the GDP) in the European Union by 2060 even by assuming no shift between the modes of care and constant prices (Lipszyc et al., 2012). To relate this prediction to the German case and its aging population, the second line of Figure 1b (80+ dependency ratio) represents the fraction of people above and including age 80 also related to people aged between 15

and 64. Since the 1950s, this fraction climbed from 1.8 to 9.3 percent of the population. It better reflects the fiscal strain for the health and especially the LTC system (as opposed to the pension funds) since the LTC costs continuously increase in age (pensions are paid discontinuously at the age-cutoff). To single out this point, the faint dashed line plots the development of the share of octogenarians and older to sixty-year olds and older. As is visible from that statistic, the share increased. The fact that it will remain on a high level in the future highlights the high demand for age-related care in the years to come. This makes a clear point: the demand for LTC services is among the severe and instantaneous consequences of the demographic transition.

The educational expansion

Apart from being a major change to all affected individuals within a society, the educational expansion closely interferes with the demographic transition. Education changes all parts of life of individuals: their job, their wage, their circle of friends, and potentially even their partner. For a first approach to the consequences of education, it is informative to look at the evolution of the educational attainment rates on the macro level. As is visible from Figure 2, a rising number of individuals received higher education. I define higher education to comprise university education (which is used interchangeably with college education) and academic track education (the highest track in the German second education tracking system and, simultaneously, the prerequisite for university enrollment).



(a) Upsurge in higher educational attainment rates

(b) Real expenditures on education per capita in West Germany (in €, deflated to year 2000)

Figure 2: Descriptives of the educational expansion in Germany

Notes: The evolution of the higher educational attainment rates in the German population are taken from Statistisches Bundesamt (2014). After 1990, the data includes East and West Germany. Reunification (vertically-dashed line) may thus confound a clear trend, but the general increasing trend remains clear. The shares relate to all individuals aged 15 and above. The per capita expenditures on education are calculated from official statistics on total real expenditures taken from Franzmann (2006).

From 1976, the earliest year with available data, around 2.3 million people had a university diploma in Germany. This number increased nearly fourfold to reach 9.5 by 2013. Although the German Reunification may confound and overstate the trend for West-Germany after 1990 (due to a converging educational attainment rate in East Germany), these figures nonetheless reflect a meaningful trend for Germany as a

whole. To better grasp a trend in the population, the dotted line with diamonds plots the trend in the corresponding share of university graduates in the population. This share increased from 3.2 in 1976 to 8.3 percent in 2013. As regards academic track attendance, the increase is even higher also in relative terms. In 1976, only 3.6 million people in the German population had graduated from the academic track. This number increased by more than the factor of five to reach 19.6 million academic track graduates in 2013. The corresponding population share nearly coincides: it increased from 7.7 to 27.9 percent in that time frame. All in all, these numbers reflect the massively increased importance of education from the macro perspective. This increased importance calls for an analysis of the causes and consequences on the individual level.

The prerequisites and the roots for the boom in higher education were, however, set before. To see the inception of this boom, it is informative to consider the annual per-capita expenditures on education in West Germany form 1950 to 1989 (the year before the Reunification). This statistic is shown in the right panel of Figure 2 for the academic track as well as for university education. The per-capita expenditures in real terms (inflated to the year 2000) on academic track education increased from € 25.0 in 1951 after the Federal Republic was founded to € 160.6 just before the Reunification in 1989 – a more than sixfold increase. Shifting from secondary to tertiary education, the expansion of the educational system becomes even more apparent. Whereas in 1951 each citizen indirectly paid on average € 18.2 for university education, this number went up to € 411.7 in 1989 – a twenty-twofold increase. This elucidates on the substantially increased importance of education in the German post-war society. Education matters to society, therefore its consequences are of public interest. Nonetheless, it is surprising how little is known about the causal individual consequences of education (Barrow and Malamud, 2015). To shed more light on this general question, Figure 2 also hints at the natural experiment that is exploited in this thesis, which came about because of the abrupt increase in the expenditures on college education: the massive extension of college spots, partly via the opening of new universities.

Data and econometric methods

Every single study of this dissertation uses empirical microdata and microeconometric estimation methods. Details on both are outlined in the following.

Data

The utilized data sources are diverse and each have both advantages and disadvantages. Yet, the data employed for each chapter-specific research question are suitable and appropriate for finding novel evidence, which complements the existing literature.

To test the intergenerational transmission of health in Chapter 1, data on multiple generations are needed. This kind of data that additionally covers the whole lifespan of the second generation to allow for a comprehensive picture on mortality differentials is usually scarce in the literature. This chapter makes use of a unique,

purpose-built administrative dataset on individuals born in Sweden between 1930–34 and their parents and on various health and socioeconomic outcomes. In total, it includes 25,064 second-generation births. From the death of this generation back to the births of their mothers, it covers a time span of 130 years (see Bhalotra et al., 2017 for a related study). Among the observed indicators for the parent generation are the date and place of birth, the socioeconomic and marital status as well as the number of children. Information on the second generation include, e.g., the sex, the place of birth, mortality and its causes, and socioeconomic outcomes in adulthood. All data are taken from official registers, such as church books and the mortality database. The latter has exceptionally low attrition since it also includes mortality information on emigrants.

For the Chapters 2 and 3, the Socio-economic panel (SOEP) is used. The SOEP is an annually-repeated household panel study on about 22,000 individuals on nearly all aspects of life (Wagner et al., 2007). The great advantage of the SOEP is its long panel dimension and the richness of the information available. Both may make a "selection-on-observables" approach (using a conventional regression or more sophisticated methods of matching) credible. Additionally, the panel dimension allows for a comprehensive estimation of the dynamic effects over time (which distinguishes the SOEP from comparable datasets as the SHARE, which has less waves and only covers individuals aged 50 and above; see Börsch-Supan and Jürges, 2005). The data set is very appropriate for the analysis on informal care, because since 2001 it includes a nonrestrictive question on the exact daily hours of informal care provision (it is not asked for the hours of care provision in the SHARE).

The National Educational Panel Study (NEPS) is used for the last three chapters on education. The NEPS is a multi-cohort panel study that consists of six different starting cohorts at different ages and positions in the German educational system (Blossfeld et al., 2011a). Each cohort is followed over time to cover their educational trajectories in detail as well as the evolution of (cognitive) skills. For Chapter 4 and 5, the adult starting cohort is employed, which comprises about 17,000 individuals born between 1944 and 1989. Chapter 6 employs both starting cohorts in the secondary education system. The cohort that attends grade 5 at the beginning of the survey comprises 9,622 students, while the cohort from grade 9 includes 16,425 students. Beside the explicit focus on educational decisions, the comparative advantage of the NEPS (with respect to any other available multipurpose household panel survey in Germany, such as the SOEP) is mainly twofold. On the one hand the NEPS provides information on regional mobility of individuals as they move through the educational system. For instance, it covers not only the district of birth but also the district of high school graduation (whereas the SOEP only includes the district of residence). These information are pivotal for the identification strategy of Chapters 4 and 5. On the other hand, the NEPS includes information on students and their teachers for the cohorts that attend elementary, primary, or secondary education. In combination with the fine-grained and objective assessment of skills and test scores, the NEPS is highly suitable and appears almost tailor-made also and especially for the research question of Chapter 6.

Empirical methods

For all chapters, causal evidence is necessary for giving credible policy advice. On

the downside, however, all chapters suffer from endogeneity problems that impede causal interpretation if they are not taken into account. This stems, for instance, from the fact that caregivers and non-caregivers or individuals with and without college education usually differ in more respects than just the considered "treatment". This difference may even go beyond what the data captures, which makes techniques that also control for "selection-on-unobservables" indispensable. Therefore, all chapters require the application of different econometric techniques that aim at isolating the ceteris paribus variation of the treatment and thus recover causal parameters. These methods then are adjusted to the respective peculiarities of the research question, the institutional setting, or the employed data.

For instance, in Chapter 1, a rather straightforward fixed effects setup is specified that, however, includes many fixed effects dimensions on a small-scale and region-specific trends. Thereby, the model captures any persistent differences across space and time in Sweden. Moreover, multiple linear-probability models are estimated subsequently that may trace out non-linear effects across the life-cycle of the second generation – the effects of interest.

Chapter 2 employs a propensity score matching approach that additionally regression-adjusts any potentially remaining differences. As control variables serve a large set of features that control for the general necessary prerequisite that someone close becomes care dependent. Moreover, the willingness and the physical or mental capability to provide care are employed, including the pre-treatment outcome to capture individual fixed effects. Additionally, a sensitivity analysis is estimated (Nannicini, 2007) that is supposed to make the identifying "selection-on-observables" assumption more credible. Chapter 3 is very similar to the preceding chapter and also uses the sensitivity analysis but refines this approach in two respects. First, dynamic matching techniques are used that allow to estimate and disentangle the effects from different paths arising from dynamic care decisions (it thus can differentiate between effect of three consecutive periods cared versus, for instance, only one). Second, it incorporates a more flexible model/variable selection for the propensity score estimation. By using a "post-double-selection" method (by employing a least-absolute-shrinkage algorithm, called "lasso", see Belloni et al., 2014), observable factors are selected more flexibly to control rather non-parametrically on important confounding factors.

Chapter 4 and 5 use different types of instrumental variables (IV) estimators that are capable of identifying causal relationships also if there are relevant factors that are unobserved. Specifically, the supply-driven extension of the college landscape in Germany (which embraces the opening of 27 new universities) is exploited as a natural experiment that delivers exogenous variation, which, in turn, is boiled down to one specific and meaningful instrument of a high statistical power. While Chapter 5 employs a rather conventional two-stage least squares (2SLS) setup together with a complier analysis, Chapter 4 estimates marginal treatment effects (MTE) that allow to unfold the effects by the unobserved taste (or preferences) for college education (Heckman and Vytlacil, 2005). This dimension of an effect heterogeneity is informative for mainly two reasons. First, it helps to recover treatment effects that go beyond the average effect of those who react to the instrument, which increases the external validity. Second, this effect heterogeneity is able to detect selection patterns that,

for instance, may answer whether individuals with the greatest taste for college education benefit most. Taken together, both points may help to assess more policy relevant effects.

The remaining Chapter 6 utilizes a difference-in-differences setup by exploiting between-subject variation in students' test scores and their teachers. By this, it can distinguish also confounding school sorting effects in addition to school, class or general time effects, of effects due to the teachers' experience. One feature exemplifies the adjustment of conventional methods to the specific needs of the research question. In order to interpret the results of the difference-in-differences model, place them into the literature, and to back up the specification of the empirical model, I set up a simple theoretical model. This model shows how the average quality of a specific cohort of teachers is affected by some marginal teachers. The specification that I derive by this model mechanically adjusts the effect to the marginal teachers. This feature emphasizes the equivalence to the IV method.

Summary of the six studies

The first Chapter (**"The long shadows of past insults – intergenerational transmission of health over 130 years"**, joint work with Claudia Andreella, Martin Karlsson, and Therese Nilsson) investigates the intergenerational transmission (IGT) of health in the very long run. Using a unique administrative dataset on individuals born in Sweden between 1930–34 and their parents, we study the intergenerational transmission of health and the impact of previous generations' health shocks on socioeconomic outcomes. As such health shocks may have serve differences in the infectious disease environment. After having demonstrated that the infant mortality rate is largely driven by infectious diseases at the turn of the last century in Sweden, short-term fluctuations in the local infant mortality rate are used to capture the disease environment. Our results provide strong evidence in favor of IGT of health, in particular for males. However, the story appears to be complex and multifaceted: the IGT exhibits an inverted socioeconomic gradient meaning that second generation individuals from a higher socioeconomic background exhibit higher effects than individuals from adverse backgrounds. This study extends the existing literature mainly by shifting the perspective to the complete life course (as opposed to, for instance, Almond et al., 2012) while exploiting less extreme health shocks than those caused by rare and devastating events.

The subsequent two chapters deal with the indirect costs of informal care provision using data from the German Socio-economic Panel. Chapter 2 (**"Short- and medium-term effects of informal care provision on female caregivers' health"**, joint work with Hendrik Schmitz) analyzes the mental and physical health effects that may be caused by informal caregiving. In contrast to the existing literature, the effects are assessed up to seven years after care provision is started. This dynamic perspective is important to completely assess the hidden costs of informal care. For instance, health effects may be persistently negative independent of whether the individual continues to provide care or stops care giving. The findings implicate that there is a considerable negative short-term effect of informal care provision on mental health, which fades out over time. Five years after care provision the effect is still negative but smaller and insignificant. Both short- and medium-term effects on physical health are virtually zero throughout. These effects are identified by exploiting the panel structure of the data that allows to control for, e.g., persistent health that is correlated with the decision to provide care (so-called reverse causality). To scrutinize the identifying assumption that all relevant variables for both health and providing informal care are controlled for, a simulation analysis assesses the sensitivity of the results with respect to potential deviations from the conditional independence assumption (necessary for "selection-on-observabels" approaches) in the regression adjusted matching approach.

Chapter 3 (**"Informal care and long-term labor market outcomes"**, joint work with Hendrik Schmitz) rounds off the evidence from the second chapter on the hidden costs of informal care provision by looking at more direct fiscal costs of care: foregone taxes that manifest through reduced labor supply and/or a wage penalty through missed promotions or atrophy in specific skills. In other words, this study presents and discusses long-run estimates on effects of informal care provision on female caregivers' labor market outcomes up to eight years after care provision. The static

version (equivalent to the empirical model of the preceding chapter), where average effects of care provision in a certain year on later labor market outcomes are estimated, is complemented by a partly dynamic version where the effects of up to three consecutive years of care provision are analyzed. The results indicate significant initial negative effects of informal care provision on the probability to work full-time. The reduction in the probability to work full-time by 4 percentage points is persistent over the course of the following eight years. Short-run effects on hourly wages are zero but amplify to considerable long-run wage penalties. The results are corroborated by a partial identification (equivalent to the sensitivity analyses of the second chapter) effect of 2.4 to 5.0 percentage points that manifest if relevant unobserved factors of certain degree still exist.

Both studies mainly contribute to the literature by shifting the focus from simultaneous or very short term effects of care on health or labor supply to dynamic effects that go beyond one or two periods after care provision (with the exception of Fevang et al., 2012 and – in the meantime – also Heger and Korfhage, 2017 who, however, looks at shorter time periods). Another contribution relates to the identification of the effects. Existing studies relied either on less-flexible regression models that relied on an uncompromising selection-on-observables assumption or used instrumental variables approaches whose results are indisputable neither. Thus, the studies of these chapters extend the literature by using highly-flexible state-of-the-art econometric techniques that also provide meaningful estimates for certain deviations from the identifying assumptions.

The remaining three chapters are dedicated to the consequences of the educational expansion. Chapter 4 and Chapter 5 are directly at the nexus of demographic change and educational expansion. Chapter 4 (**"Heterogeneity in marginal non-monetary returns to higher education"**, joint work with Daniel Kamhöfer and Hendrik Schmitz) starts by estimating the effects of college education on cognitive abilities, health, and wages. The effects of college education are identified by using an arguably exogenous variation induced through college expansions (the opening of new universities and the increase in the capacity of existing ones). In addition, semiparametric local instrumental variables techniques are applied that allow to identify marginal treatment effects in an environment of essential heterogeneity. The results suggest positive average effects on cognitive abilities, wages, and physical health. Yet, there is heterogeneity in the effects, which points toward selection into gains. While the majority of individuals benefits from more education, the average causal effect for individuals with the lowest unobserved desire to study is zero for all outcomes. Mental health effects, however, are absent for the entire population.

Chapter 5 (**"Fertility effects of college education: evidence from the German educational expansion"**, joint work with Daniel Kamhöfer) assesses female fertility patterns while using the same instrument as in the previous chapter. However, also with respect to female fertility, college education is a so far understudied margin of education: no study aims at identifying of causal effects while explicitly focusing on fertility (Currie and Moretti, 2003 consider fertility implicitly as a potential channel) – the apparent contribution of this study. While college education reduces the probability of becoming a mother, college-educated mothers have slightly more children than mothers without a college education. Unfolding the effects by the timing of birth reveals a postponement that goes beyond the time in college – indicating a negative

early-career effect on fertility. Coupled with higher labor-supply and wage returns for non-mothers (as compared to mothers), the timing of the effects moreover suggest that career and family are not fully compatible.

The contribution of both studies are straightforward, since – opposed to anecdotal evidence – surprisingly little is known about the causal effects of college education even for monetary returns (Barrow and Malamud, 2015). Evidence is even more scarce when it comes to the considered non-monetary outcomes. Neither has the literature provided any long-term effects of college education on cognitive skills, nor delivered a comprehensive study on female fertility effects. In addition, the contribution is extended by employing state-of-the art econometric methods that additionally allow to unfold the effects, for instance, by the timing of the births: both allow for novel insights into the respective research question (heterogeneity on selection patterns for the marginal effects and a faint understanding of the underlying mechanism for the fertility effects).

Chapter 6 (**"More teachers, smarter students? Potential side effects of the German educational expansion"**, single-authored) evaluates the potential side effects of the educational expansion in Germany on the learning outcomes of today's students. The educational expansion was a demand shock in the labor market of teachers, which could have thus encouraged individuals with different teaching abilities to eventually become teachers. I find that replacing a non-affected teacher with an educational expansion teacher leads to a 2 percent reduction in students' test scores. Explorative analyses suggest that the educational expansion teachers are more extrinsically rather than intrinsically motivated. Furthermore, these teachers also exhibit worse grades in their highschool exit exams. The results generally highlight that monitoring and investing in quality (of, for instance, the vocational training) is important for future extensions of public institutions.

This study adds to the literature by assessing explicitly educational expansion-induced repercussions in the labor market of teachers, while placing the findings into two strands of closely-related studies. First, the causal effects of teachers on student test scores (generally identified via teacher fixed effects, see e.g. Hanushek, 1971, Hanushek and Woessmann, 2008 or Chetty et al., 2014a) and its long-term effects on students' life-time earnings (Chetty et al., 2014b). And second, studies on the selection of individuals into the teaching profession (see, e.g. Britton and Propper, 2016 or Nagler et al., 2015). Furthermore, this study may generate novel evidence by going beyond the US and looking at students that are much more homogeneous in skills (as they attend the academic track in Germany).

Part II

Consequences of the demographic transition

Chapter 1

The long shadows of past insults – intergenerational transmission of health over 130 years[1]

1.1 Introduction

The origins of societal inequality in health and socioeconomic outcomes are not well understood – but their persistence over time is quite remarkable. A large body of literature in economics measures intergenerational correlations in socioeconomic status. In the U.S., the elasticity of sons' earnings with respect to their fathers' earnings is 0.5-0.6 (Mazumder, 2005); whereas the corresponding number for Sweden is 0.25-0.3 (Jantti et al., 2006; Lindahl et al., 2012). Similar results are found for other socioeconomic outcomes such as IQ and educational attainment (Hertz et al., 2007). A nascent strand of the literature suggests that the persistence in social hierarchies may be even stronger than what a narrow focus on labour market outcomes would suggest (Clark, 2012).

A parallel, but considerably smaller, literature in health economics studies the intergenerational transmission of health. A typical indicator is birth weight, which has been shown to be persistent across generations. For American twins, Royer (2009) finds that a 100-gram increase in birth weight associates with an 18-gram increase in the following generation; estimates including mother fixed effects are smaller but strongly significant. Intergenerational persistence is also observed in other health outcomes, such as longevity and self-assessed health (Trannoy et al., 2010) or body mass index (Classen, 2010).

Despite these efforts, the literature does not deliver much evidence of the extent to which the associations are modifiable. Comparing biological and adopted children, Thompson (2014) estimates the genetic component in the intergenerational correlation of some common chronic diseases and concludes that it is surprisingly low,

[1] This chapter is written together with Claudia Andreella, Martin Karlsson, and Therese Nilsson. It is published as a working paper as: Andreella, C., Karlsson, M., Nilsson, T., and Westphal, M. (2015). The Long Shadows of Past Insults Intergenerational Transmission of Health over 130 Years. Ruhr Economic Papers 571, RWI Essen.

explaining only 20-30% of the observed intergenerational link. This result, however, still begs the question as to whether the remaining intergenerational link is modifiable. A large and growing literature documents that early life shocks may have strong effects on health and socioeconomic outcomes in adulthood (Scholte et al., 2012; Almond, 2006; Almond et al., 2009). The same appears to hold for policy interventions (Bhalotra et al., 2015; Bharadwaj et al., 2011; Bhalotra and Venkataramani, 2011; Aizer and Currie, 2014). However, it is still largely unclear whether these early life influences contribute to reducing inequalities in health and socioeconomic outcomes in later generations. Increasing our knowledge about this issue would clearly be highly desirable, since it may be the case that public interventions generate returns over a very long time period.

It is the purpose of this paper to closely examine the intergenerational transmission of an early life health shock. Using extremely detailed Swedish data on the in utero disease environment of the first generation, and on various health and socioeconomic outcomes of the first and the second generations over a time span of 130 years, we seek to answer three distinct questions. First, we estimate at what rate the impact of the initial health shock diminishes from one generation to the next. Second, we analyse whether there is a socioeconomic gradient in the intergenerational transmission of health. Third, we estimate the repercussions of the initial health insult on a variety of socioeconomic outcomes in the second generation. We contribute to the existing literature by shifting the perspective to the complete life course (as opposed to Almond et al., 2012) while exploiting less extreme health shocks than those caused by rare and devastating events (Richter and Robling, 2013; Van den Berg et al., 2014). The first generation in our dataset was born in the decades around the turn of the 20th century – after the famine of 1866–8 (Doblhammer et al., 2011) and the last outbreak of smallpox in 1873–4 (Sköld, 1996), but before the Spanish flu pandemic in 1918 (Karlsson et al., 2014) – in an era characterised by gradual improvements in public health and by the absence of severe mortality crises.

Our analysis makes use of a unique and purpose-built Swedish dataset. A general challenge for an empirical analysis of this kind, even in Scandinavia, is that there are hardly any datasets which cover the long time spans and variables required without serious attrition due to mortality and migration.[2] We thus mainly rely on tailor-made datasets and match them to administrative data whenever possible. Our main dataset is a representative sample of 25,000 births from the cohorts 1930–34. It includes a large number of indicators for the parent generation – e.g. date and place of birth, socioeconomic and marital status, number of children – as well as a wide range of outcomes for the second generation – e.g. sex, place of birth, mortality, and socioeconomic outcomes in adulthood. All data are taken from official registers and the mortality database has exceptionally low attrition since it also includes mortality information on emigrants.

[2]For example, the Swedish multigeneration register contains only individuals born after 1932 who were alive and registered at some point after 1960. Information on parents is complete only from the 1950 cohort onwards (Statistics Sweden, 2005). Moreover, existing demographic intergenerational databases (in Sweden, Canada, Italy, Switzerland and other countries) generally cover complete regions (e.g. a geographical cluster of parishes or a city) but are not representative of the entire population since the selection of areas do generally not take such criteria into account (Edvinsson, 2000). A main obstacle with existing demographic databases is also the lack of digitised individual level data for the period 1900-1950 (Bengtsson and van Poppel, 2011).

Our indicator on the first generation disease environment is the local infant mortality rate in the parish of birth of the mother. For 1,856 different birth parishes, we compiled information on annual IMR for the period 1880-1917. This variable is commonly used as a proxy for disease environment in utero and early childhood, and has been found to be an important predictor of adult health (Akachi and Canning, 2007; Bozzoli et al., 2009; Crimmins and Finch, 2006) and various other outcomes (Lawson and Spears, 2014; Case and Paxson, 2009). Some authors argue that post-neonatal mortality (PNM) is a better indicator of early life disease environment since it does not include neonatal mortality, which is also strongly associated with the access to pre- and perinatal care (Schmidt et al., 1995). PNM rates are not available for the time period we consider, but the drawback is probably less of a concern when considering conditions in Sweden in the late 19th and early 20th centuries, since access to those services was limited (Bhalotra et al., 2015) and their efficacy in improving infant health has been challenged (Pettersson-Lidbom, 2014). In auxiliary regressions, we confirm empirically that infectious disease is a much more important predictor of the IMR than access to health care. Nevertheless, we include a large battery of local fixed effects and trends in order to reduce the influence of persistence in local differences in public services and other confounding factors. Consequently the identifying variation we exploit are deviations of the IMR from local trends and levels.

A more serious concern regarding the use of the local infant mortality rate as an indicator of disease environment is the issue of selection (Almond et al., 2012). In particular if identification is driven by large deviations from the local mean, one should be concerned that the resulting sample will be very strongly selected (cf. Doblhammer et al., 2011). Any impact of local IMR on later-life outcomes would then be a combination of scarring and selection effects operating in opposite directions. Using a setting similar to ours, Hatton (2011) finds no evidence of selection, and Bozzoli et al. (2009) conclude that it may be more of an issue in developing than in developed countries. For the cohorts we study, the national IMR dropped from 10 per cent to 6 per cent during the observation period. This would be high by today's standards in developing countries, but much lower than the levels that have been experienced in developing countries in the past. However, it is very clear in our case that scarring dominates selection: the mothers who suffered an unfavourable disease environment had elevated mortality at older ages, and their children experienced worse health and SES outcomes.

In general, our results corroborate earlier studies on the importance of the in utero environment: we find that a maternal health shock affects survival prospects in the second generation, that this effect is small or even non-existent during the first decades of life, but that it has a clear and significant impact on survival prospects after the age of 50 (Almond and Currie, 2011). The effect is also particularly pronounced for males: if the local IMR in the place of birth of the mother increases by ten percentage points, the risk of dying before the age of 70 increases by three percentage points (ten per cent) within this group. We do not find any evidence of this effect being driven by fertility responses.

In two respects our results however differ strongly from the typical findings of the previous literature. First, we find evidence of an inverse SES gradient in the intergenerational transmission of health. When the survival disadvantage becomes visible in the second generation from age 50 onwards, it is strongly concentrated among

individuals with a better-than-average SES background. In particular males from a privileged background seem to be affected. This is in stark contrast with the previous literature, which almost always finds that the intergenerational transmission is stronger in disadvantaged groups (Bhalotra and Rawlings, 2013; Currie and Moretti, 2007; Kim et al., 2014; Costa-Font and Gil, 2013).[3] By considering grandparental SES we are able to rule out the possibility that our results are driven by selection into the advantaged group. Instead, we provide evidence suggesting that this unusual result is attributable to environmental and behavioural factors in childhood and adulthood: looking at specific death causes, related in particular to cardiovascular diseases and diabetes, there is compelling evidence that in utero metabolic adaptations have taken place. Such changes are compatible with the thrifty phenotype hypothesis (Hales and Barker, 1992) and may be particularly likely to lead to metabolic disorders in an affluent environment (Barker, 1997). This candidate explanation is corroborated by surveys of dietary habits in Sweden in the years following the births of our second generation (Boalt, 1939; Odin, 1934). Besides, we find evidence suggesting that some of the gradient is driven by behavioural changes in adulthood.

Second, our results regarding the impact on labour market outcomes are equally intriguing. For earnings, we find a strong impact of the maternal disease environment: a one-point increase in maternal IMR is associated with a one-percent reduction in adult earnings. The effect is large: it is comparable to the impact found within the first generation by Lawson and Spears (2014) or to the effect of a ten-percent increase in birth weight (Black et al., 2007). Surprisingly, the effect is entirely driven by females. This result contrasts the findings in some previous literature which suggests that health shocks disproportionately affect the labour market outcomes for males (Black et al., 2007; Cai, 2010; Pelkowski and Berger, 2004). In our case, men's labour market outcomes are hardly affected, but elasticity of female earnings is -2 – a one percentage point increase in the maternal IMR is associated with a 2 per cent reduction of second generation female earnings in the 1970's. The effect appears to be particularly strong for females from low-SES backgrounds. We are able to show that this effect is linked to the expansion of the welfare state: for the relevant cohorts, female employment increased in particular in public services, and there appears to have been positive health-related selection into these professions.

The disadvantage associated with a maternal health shock is thus complex and multifaceted. But at what age is the disadvantage determined? Analysing educational attainment in the second generation, we find weak evidence that the disadvantage is fixed early in life: the maternal disease environment associates with shorter completed education for high-SES males and low-SES females – i.e. the groups that appear to be particularly affected – but the coefficients are generally small and not statistically significant. On the other hand, our analysis by death cause delivers relatively

[3] Almond et al. (2012) do find a similarly inverted black-white gradient, which they attribute to selection effects dominating amongst black mothers. This explanation is unlikely to apply in our case, since we find evidence of scarring in the first generation also for the disadvantaged group.

strong evidence of an epigenetic transmission. The recent epidemiological literature points to a relationship between "viral infections and preterm labour, and fetal congenital anomalies of the central nervous system and the cardiovascular system"(Mor and Cardenas, 2010).[4] According to the epigenetics story, environmental conditions in utero and early life may alter the genetic phenotype: for instance, if the disease load is high, bacteria could cross the placenta causing an inflammatory response syndrome which can have long-term consequences. There is some evidence that such phenotypic changes can also be transmitted across generations Hochberg et al. (2011).

The cohorts we study are special in the sense that they were born around the Great Depression, which evidently represents an environmental shock to the second generation. However, we do not find any evidence of the Great Depression having a detrimental effect on health and labour market outcomes of the cohorts exposed to it in early childhood. Using administrative data of the crisis impact at the local level, our difference-in-difference estimates for adult incomes are insignificant throughout. As regards health, we find evidence that the crisis had a protective effect on mortality from age 50 onwards. This finding is, on the one hand, consistent with the common result that downturns are associated with improvements in health (Ruhm, 2000, 2003, 2004; Neumayer, 2004; Gerdtham and Ruhm, 2006) but on the other hand deviates from the literature focusing on the early life enviroment, which generally finds adverse effects (Van den Berg et al., 2006; Lindeboom et al., 2010; Van den Berg et al., 2009, 2011). We leave the investigation of this result as a topic for later research, but conclude that the effects of the crisis seem to apply across the board since we find no evidence that the crisis moderates the effects of the maternal health shock.

1.2 Literature review

Studies exploring the Barker hypothesis can be broadly classified according to the timing of the outcome variable: either studies limit their assessment of intrauterine shocks to one generation (early life or in adulthood), or they extend their scope on the intergenerational transmission mechanism.

Studies that focus on intragenerational effects further differ with respect to the health shock used to proxy the environment experienced by individuals. Such studies examine health shocks that either affect the fetus while in utero, or hit the individual very early in life (at birth or up to the first years of life). This distinction is important for identification of the underlying mechanism. Moreover, postnatal measures are more likely to be under the control of and are thus either more likely to be compensated by parental investment or correlated with unobserved parental characteristics.

Birth weight is the most widely used proxy for health at birth. However, studies using this proxy have the potential common deficiency that birth weight is not determined independently of maternal behavior, resulting in estimates that might not have a causal interpretation (see e.g. Coutinho et al., 1997; Currie and Moretti, 2007; Datar

[4] But also maternal stress has found to be a driver of preterm labour and, thus, of infant mortality (Goldenberg et al., 2008).

et al., 2010; Royer, 2009, for a discussion). Birth weight might e.g. be influenced by lifestyle during pregnancy, which is unobserved and thus a threat to identification.[5] As Currie (2009, p.102) states: "the most compelling examinations of the fetal origins hypothesis look for sharp exogenous shocks in fetal health that are caused by conditions outside the control of the mother".

The literature points at two types of exogenous shocks, entailing that the Barker hypothesis may work through different potential mechanisms. First, the literature has evaluated direct health shocks that hit a given area. Examples of surveys that exploit such exogenous shocks are the "Dutch Hunger Winter" analyzed inter alia by Scholte et al. (2012), huge pandemics like the devastating Spanish flu in 1918 (Almond, 2006) and the nuclear disaster of Chernobyl (Almond et al., 2009). Such shocks can be seen as good sources of exogenous variation in the health environment. Another widely used proxy for the health environment at birth or in utero is the infant mortality rate. Although its effect might be theoretically ambiguous, as only surviving infants are considered, it has been shown to be a good proxy for early health environment, because it accounts for huge differences in the cross-cohort health outcomes (mortality and height; see: Bozzoli et al., 2007, Crimmins and Finch, 2006)

Second, some papers explore the role of external economic shocks. However, as pointed out e. g. by Goldstein (2013) on financial crisis, such shocks may be foreseeable to some extent. If it is the case, they can influence fertility decision of hypothetical parents. Beyond that, the exact transmission channel is ambiguous. The main channels proposed in the literature are the level of public health expenditures, the change in the opportunity cost of health care, nutritional deficiencies due to the loss of financial resources and psychosocial stress. We focus on the last two mechanisms, as they are linked to in-utero exposure (see for instance Margerison-Zilko et al., 2011). Although deficiencies in specific nutrients could indeed influence mortality, there is no consistent evidence showing that, in our setting, it would actually be the case. Therefore, we are more inclined to focus on the psychosocial stress channel as the main transmission mechanism between economic and health shocks (Bejenariu and Mitrut, 2013).

Only a few studies focus the intergenerational effect of exogenous health shocks. Drake (2004) was the first to extend the scope of the fetal-origins hypothesis to subsequent generations. They evaluate the fetal origins hypothesis in terms of the second generation's birth weight and find some evidence of non-genetic inheritance. One of the mechanisms through which this effect operates is epigenetics. According to epigenetics, environmental conditions in utero and early life may alter the genetic phenotype. This can be efficient from an evolutionary point of view since it means that the fetal metabolism reacts to outside information and is hereby adjusted to the environment it is expected to be born into.[6] These phenotypic changes can also be inherited by the next generation (Jirtle and Skinner, 2007).

[5] However, twin differences in birth weight fix these confounding factors but also hold the intrauterine environment constant (see, for example, Figlio et al., 2014).

[6] For example, Prentice (2006) review literature that builds the bridge between maternal malnutrition and offspring's obesity. The main explanation is that those fetuses had to use resources more efficiently which makes them more susceptible to gain weight in later life.

The epidemiological literature has shown a relationship between "viral infections and preterm labor, and fetal congenital anomalies of the central nervous system and the cardiovascular system"(Mor and Cardenas, 2010)[7]; if the disease load is high, bacteria could cross the placenta causing an inflammatory response syndrome which can have long-term consequences. Even today, preterm births account for 75% of perinatal mortality and they are often caused by intrauterine infections (25-40% of the cases, Goldenberg et al. (2008)).

Bhalotra and Rawlings (2013) is among the few studies that exploits exogenous variation in several measures related to economic growth in developing countries in order to analyze the gradient in the intergenerational transmission of health. By controlling, inter alia, for family specific endowments, they find that a change in the health environment of the first generation has some effects on early-life health outcomes of the second generation. We contribute to this literature by considering health outcomes also later in life. In addition, we examine the heterogeneity of such effects, driven by sex and by the transmitted socio-economic status in the second generation.

1.3 Data

One of the reasons why estimates of the intergenerational transmission of health and of the intergenerational effect of health on socio-economic outcomes are scarce is that data pose a significant challenge. Such an analysis requires tracking a sufficiently large number of individuals over the life-course and at the same time have information on the birth location and early-life disease environment of the parental generation. Other general challenges are attrition, mortality in early life and the traceability of migrants. The following two sections describe how we construct a tailor-made dataset which adresses most of these concerns.

1.3.1 Individual level data

The base component of our dataset is a representative sample of individuals born in Sweden from January 1, 1930 to December 31, 1934, whom we follow over the life course. We construct the micro-level data by digitising parish records from 133 parishes throughout the country on all individual births, including detailed information on birth date, place of birth, name, sex, mother's marital status, parents' name, date of birth and occupation.[8]

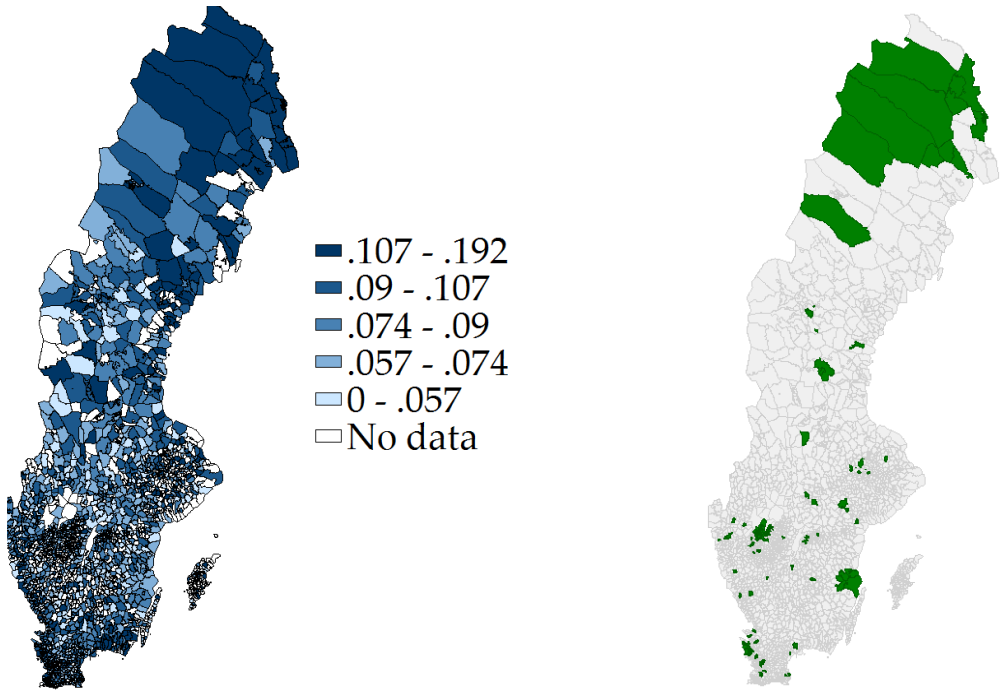
The Swedish church law of 1686 states that the clergyman in each parish should keep record for all children born in and out of wedlock. Vital statistics were thus introduced very early in Sweden and the information provided by parish records are generally seen as being of very high quality (Edvinsson, 2000). Covering everyone born during the sample period in the selected parishes the data includes 25,064 individual

[7] But also maternal stress has found to be a driver of preterm labor and, thus, of infant mortality (Goldenberg et al., 2008).

[8] Parishes (församling/socken) are subdivisions within the Church of Sweden. In 1930s there were about 2,200 parishes in the country.

births corresponding to nearly 6 per cent of all births in Sweden in 1930-1934. The data also allow us to identify siblings born within this time period and twin births. There are 12,015 siblings in the dataset, from 5,279 mothers.

Figure 1.1 shows the geographical location of the birth parishes of our individuals. Birth parishes are distributed across all parts of the country and using information from the 1930 census we confirm that the locations covered in the dataset are representative of Sweden as a whole in terms of observable characteristics such as economic structure, average income and infant mortality, corroborating the external validity of our results.



(a) Average IMR between 1880 and 1919. (b) The 133 second-generation birth parishes.

Figure 1.1: Maps of first- and second-generation birth parishes.

We follow each individual from birth to death, or to age 79-83 if alive in 2013. Using name, gender and date and place of birth we trace the date of death of deceased individuals from the Swedish deathbook (Sveriges dödsbok). The deathbook contains the universe of deaths occurring in Sweden and to Swedes abroad during the period 1901-2013. In order to validate our matches we use information from cemetery burial records (Federation of Swedish Genealogical Societies, 2009).[9] As a second validation source we use tax records on annual incomes (labor, capital and business income) 2001–12 and the 1970 census.[10]

With the above sources we track 96 per cent of all individuals for a period of 79 years.[11] Among our cohorts 44.7 per cent did not survive to the age of 79. For deaths occurring after 1960 we also have information on individual death causes from the Death

[9] This source covers all grave registers in the country and includes more than 5 million entries.

[10] These sources also allow us to cross-check names, of particular importance for the correct tracing of women who often change their last name when getting married.

[11] The remaining 4 per cent are likely to be surviving emigrants; we thus include them as survivors in our analyses of mortality. Results are insensitive to the exclusion of this group.

cause register (The National Board of Health & Welfare, 2014) covering all deceased residents who died in Sweden or abroad. Looking into descriptive statistics the quality of our death records seems very high. 2.6 per cent of all births in the sample are stillbirths, the infant mortality rate is 58 and child mortality 73 per 1,000 live births. This is all to be considered normal of the 1930s (Edvinsson et al., 2008) and suggests accurate and high quality recording. Similarly the sex ratio at birth is exactly the expected 1.05. Also with respect to death cause statistics we can confirm the external validity of our dataset: 38.5 and 30.6 per cent respectively of our sample population died because of circulatory diseases and cancer in the 1990s which corresponds to official figures for the corresponding demographic groups in the entire population (Statistics Sweden, 1997).

Our health outcome variables are a mutually exclusive set of binary mortality indicators over the whole life course (censored at age 79): mortality between ages 0–1; 1–50; 50–70; and 70 and older. Even though we consider individual-level mortality from birth onwards, we put particular emphasis on mortality between 50 and 70, as these ages have been identified as a critical period in the fetal origins literature.[12]

We also examine the intergenerational transmission of health to socio-economic outcomes.[13] Our main socio-economic outcomes are measured in adult life (age 36-40) when the second-generation individuals are typically active on the labour market after having completed their education. The primary socioeconomic outcome we consider is log individual earnings from the 1970 Census.[14] We also examine other labour market outcomes for the second generation, such as employment (total, full-time, part-time) and sector of employment. Finally, we examine the impact on completed years of schooling.

Table 1.1 provides descriptive statistics for the variables in the dataset.

The parish records provide pertinent information on the parents (marital status, name, date of birth and occupation). As household SES may play an important role for health and socioeconomic outcomes of the second generation, we classify the parents' occupations according to the HISCO system (Historical International Standard Classification of Occupation) to derive relevant SES groups. The HISCO rank occupations according to the required skill level of an occupation, where 0 indicates the most skilled non-manual jobs, and 9 indicates the lowest skilled manual occupations. Table 1.2 lists the major HISCO groups, together with a short description and the absolute and relative frequencies in our data. Many household heads belong to HISCO group 6, representing mainly farmers, and to HISCO group 9, representing workers employed in unskilled jobs.[15]

To allow for a survival analysis of the first generation we also collect information on the date of death of the mothers from the death book. We identify the exact date of death of 89 per cent of our first generation population.

Digitising the parish records gives an advantage compared to official multigenerational registers since the latter only contain individuals who were alive and registered

[12] Barker's seminal work explores in particular the association between low birth weight and the incidence of certain types of diseases in middle age - among others, coronary heart disease, hypertension and diabetes (Barker, 1990; Hales and Barker, 1992).

[13] As we observe all individuals and either their income or their death (or both), we can examine the role of attrition. In particular we compare descriptive statistics of the sub-sample of individuals

Table 1.1: Descriptive statistics.

Variable	Mean	SD	Min	Max	Obs
Individual birth/death data					
IMR^j	0.10	0.05	0	1	25,010
crisis^i	0.05	0.20	-0.59	0.61	25,010
Female	0.48	0.50	0	1	25,010
Twin	0.03	0.16	0	1	25,010
Wedlock	0.89	0.31	0	1	25,005
Age mother (yrs)	29.13	6.66	13	50	25,007
Urban	0.21	0.41	0	1	25,010
Mortality 0-1 (baseline)	0.08	0.28	0	1	25,010
Mortality 1-50 (baseline)	0.08	0.27	0	1	22,940
Mortality 50-70 (baseline)	0.16	0.37	0	1	21,081
Mortality 70+ (baseline)	0.30	0.46	0	1	17,632
Census 1970					
Log labor income in 1970	8.41	3.56	0	13.37	18,566
Years of education	9.62	2.46	7.69	19	18,372
Death causes 1960-2013					
All causes	0.41	0.49	0	1	22,739
Cancer (excl. lung/oral cavity)	0.08	0.28	0	1	22,739
Lung/oral cavity cancer	0.02	0.14	0	1	22,739
Respiratory diseases	0.04	0.19	0	1	22,739
CVD (no curable risk factors)	0.09	0.28	0	1	22,739
CVD (with curable risk factors)	0.03	0.17	0	1	22,739
Other circulatory system diseases	0.03	0.18	0	1	22,739
External causes/infections	0.03	0.18	0	1	22,739
Digestive/endocrine system (incl. diabetes)	0.01	0.12	0	1	22,739
Other symptoms	0.04	0.20	0	1	22,739
Not classified elsewhere	0.03	0.17	0	1	22,739

Table 1.2: Occupation of the household head according to the HISCO classification.

HISCO cat.	Description	Freq.	Percent	Cum.
0	Professional, technical and related workers	330	1.32	1.32
1		806	3.22	4.54
2	Administrative and managerial workers	517	2.07	6.61
3	Clerical and related workers	368	1.47	8.08
4	Sales workers	637	2.55	10.63
5	Service workers	439	1.76	12.38
6	Agricultural, animal husbandry and forestry workers, fishermen, hunters	8,913	35.64	48.02
7	Production and related workers, transport	2,291	9.16	57.18
8		1,324	5.29	62.47
9	Equipment operators and labourers	6,537	26.14	88.61
10	Unknown	2,848	11.39	100
	Total	25,010	100	

who died before 1970 to the ones who did not die before 1970 and we compare our main regression results with the corresponding specification estimated using IPW (available upon request). Importantly, attrition does not seem to be an issue in our data.

[14]As discussed by e.g. Haider and Solon (2006) and Böhlmark and Lindquist (2006), income data measured at ages 30-45 generally provide a good proxy of lifetime income as they are less likely to fluctuate due to life-cycle biases.

[15]The reported information is generally based on fathers' occupation.

in Sweden at some point in time after 1960. Similarly information on the parental generation is conditioned on being alive and parental data is complete only from the 1950 cohort and onwards (Statistics Sweden, 2005). Using parish records also assure that we do not have any misreporting in the place of birth of an individual.[16] These unique features of the dataset allow us to match individual-level data for both the first and the second generation to detailed information on the local economic and in utero disease environment.

1.3.2 Maternal disease environment in utero

Our indicator of the maternal disease environment in utero is the infant mortality rate in her parish of birth. Lifetime individual-level data have previously been combined with local IMR to evaluate the impact of exogenous variation in early-life conditions of one generation (see e.g. Bengtsson and Lindström, 2003; Van den Berg et al., 2009).[17] To analyse the role of in utero disease environment of a parental generation we first use the information provided on parents in the parish birth records of the second generation and the death book to exactly identify the parish of birth of the mothers of the children of the second generation. The first generation mothers are born between 1880 and 1918 in 1,864 parishes across the country.

We calculate the IMR within the time window 1880-1918 on the parish level. For the period 1880-1900 we use IMR data from Statistiska Centralbyrån (Statistics Sweden, SCB). For the remaining years we calculate local IMR from the deathbook.[18] We obtain the IMR at birth location for 99.8 per cent of the mothers in the first generation.

Figure 1.1a shows the regional variation in the IMR level, while Figure 1.2 shows the change over time of the IMR from 1880 to 1918 averaged over all parishes.[19]

[16] In official registers the parish of birth reported for cohorts born until 1946 refers to the place of the actual birth of an individual, i.e. if an individual was born in a hospital the parish of birth reported refers to the location of the hospital in which someone was born (Skatteverket, 2015). The transition to institutional delivery started in the late 1920s and was initially very smooth, but in the mid-1940s the majority of births took place out of the home (Wisselgren, 2005)

[17] Along the same lines we use the infant mortality rate as a proxy for the health environment at birth and/or in utero. Although its effect might be theoretically ambiguous, as only surviving infants are considered, it accounts for huge differences in the cross-cohort health outcomes and it has been shown to be a good proxy for early health environment (mortality and height; see: Bozzoli et al., 2007, Crimmins and Finch, 2006). We further discuss the potential determinants of IMR in Section 1.5.

[18] The IMR for the parish of birth of the mother is missing for 16.9 per cent of the mothers in our sample. For these cases we impute IMR based on the regional annual IMR, taking the annual IMR in the region of birth and subtracting a weighted average of all available IMRs at the parish level in the same region. In addition, information on the parish of birth of the mother is missing for 14.6 percent of our first generation. In these cases we impute the IMR of the parish of birth of the firstborn child. Such imputed measures are cruder than the local IMR but improves representativeness and allows us to include additional observations.

[19] As described below we use fixed effects implying that the identifying variation in our estimation will correspond to deviations from the levels and time trends in IMR within each parish.

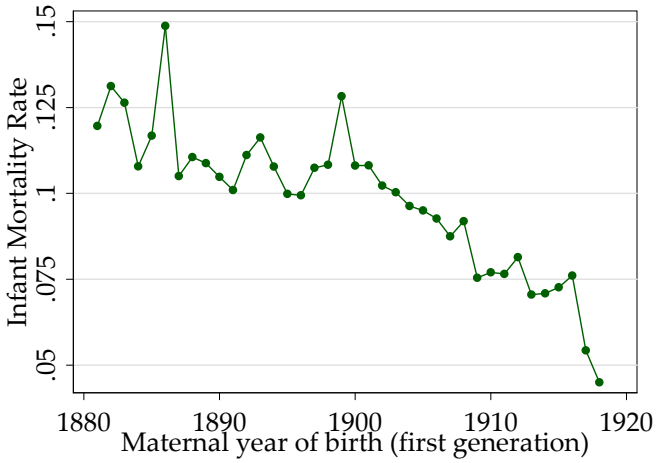


Figure 1.2: Average infant mortality rate (IMR) by maternal year of birth.

1.3.3 Economic environment

The cohorts of 1930-1934 were born around the Great Depression, which represents an additional environmental shock. The peak of the crisis in Sweden appeared in March 1932 with low per capita incomes and high unemployment (Kobrak and Mira, 2013), but the economy recovered remarkably fast: in 1934, the gross domestic product was already back to its 1930 level (Mitchell, 1998). To test if the recession had any effects on health and socioeconomic outcomes we add local level information on the general economic situation measured as deviations in annual municipality income tax revenue for the period 1930-1934, collected from yearbooks (Statistics Sweden, 1935).[20] We divide the yearly tax revenues at the municipality level by the working population in 1930 (from the 1930 Census) and we deflate this measure by the Cost of Living Indicator (CLI, by Statistics Sweden). We define the crisis indicator $crisis^{i}_{sy}$ as the negative logarithmic deviation of the deflated per capita tax revenues (tax) in year $y = (1931, \ldots, 1934)$ with respect to the same measure in 1930, for each parish s.

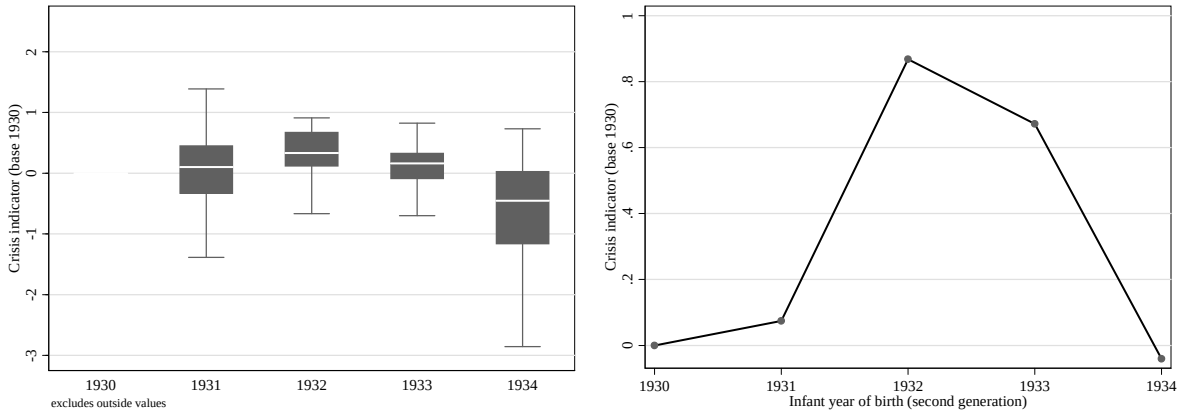
$$crisis^{i}_{sy} = -\left[log(tax_{s,y}) - log(tax_{s,1930})\right], \quad y \in \{1931, \ldots, 1934\} \tag{1.1}$$

If the indicator is positive the crisis hit a parish, and if negative the economic situation in a parish was better in year y than in 1930. Figure 1.3 shows how the indicator changes over time and across parishes. On average, the indicator increases until 1932 and then declines again until 1934.

1.4 Empirical strategy

As Figure 1.2 reveals, Sweden experienced a declining infant mortality rate over the relevant period. In addition to these temporal trends, there are also spatial patterns.

[20] At the time parishes and municipalities more or less coincided. The 133 parishes covered by our individual-level dataset are grouped in 118 municipalities.



(a) Dynamics of the distribution (b) Dynamics of the mean

Figure 1.3: Crisis indicator: average negative log deviation from 1930 tax revenues.

If these patterns are not taken into account, the IMR coefficient simply reflects the fact that mothers who were exposed to higher infant mortality rates are older (confounding time-trend) and/or were born in places where, for example, public services were scarce or had a low quality or, again, that individuals select themselves into municipalities according to their health endowment. It could also be that different regions in Sweden exhibit different rates of technological progress or expansions of the medical sector.

We control for regional fixed effects and different regional-specific trends (spatial time trends and ageing trends in the first generation) using the following baseline econometric model, similar to the one employed by Almond et al. (2012):

$$Y^{ij}_{ptrsya} = \alpha + IMR^{j}_{pt}\beta_1 + X^{i}\gamma + \theta^{j}_{a} + \theta^{j}_{a}t^{j} + \delta^{j}_{r} + \delta^{j}_{r}t^{j} + \epsilon_{ij} \tag{1.2}$$

where Y^{ij}_{ptrsya} refers to health or to socioeconomic outcomes for individual i, born in parish s in year y, belonging to the second generation. j indicates individual i's mother, belonging to the first generation, while p refers to mother j's parish of birth, t her year of birth, r the region of birth and a age when giving birth. IMR^{j}_{pt} proxies the disease environment that the individuals in the first generation experienced in utero. We use a linear probability model (LPM) to estimate Eq. 1.2 when the focus is on binary mortality indicators. For our economic indicators we estimate Eq.1.2 by OLS.

The coefficient β_1 reports the effect of the health shock, IMR^{j}_{pt}, which is the effect of interest. We argue that this coefficient is identified conditional on the covariates.[21]

[21] In Section 1.5 we discuss whether IMR is a good approximation for the health environment of the mother at birth and whether it is an appropriate measure for the effects that we are considering in the second generation.

X^i is a vector of control variables pertaining to individual i in the second generation, born in parish s in year y. It contains binary indicators for sex, whether a twin birth occurred or whether individual i is born in wedlock and dummy series for the order of birth, the quarter and year of birth, the parish of birth of individual i and the occupation of the household head. We also control for a crisis indicator to capture possible effects of the Great Depression on mortality. Furthermore we control for spatial effects on the parish level where the second generation is born and for seasonal variation.

Equation 1.2 also accounts for several characteristics of the first generation. We include maternal age fixed effects θ_a^j to account for the influence of the mother's age on the foetus' health and mother's region of birth r (Län) δ_r^j fixed effects. We control for time trends using two vectors. The first ($\theta_a^j t^j$) addresses time trends in the age of the mother. It takes into account that the age effect of mothers at birth is changing over time. The second time trend ($\delta_r^j t^j$) accounts for separate time trends by region. Holding trends and levels within regions fixed, β_1 is thus identified by deviations from those trends and levels.

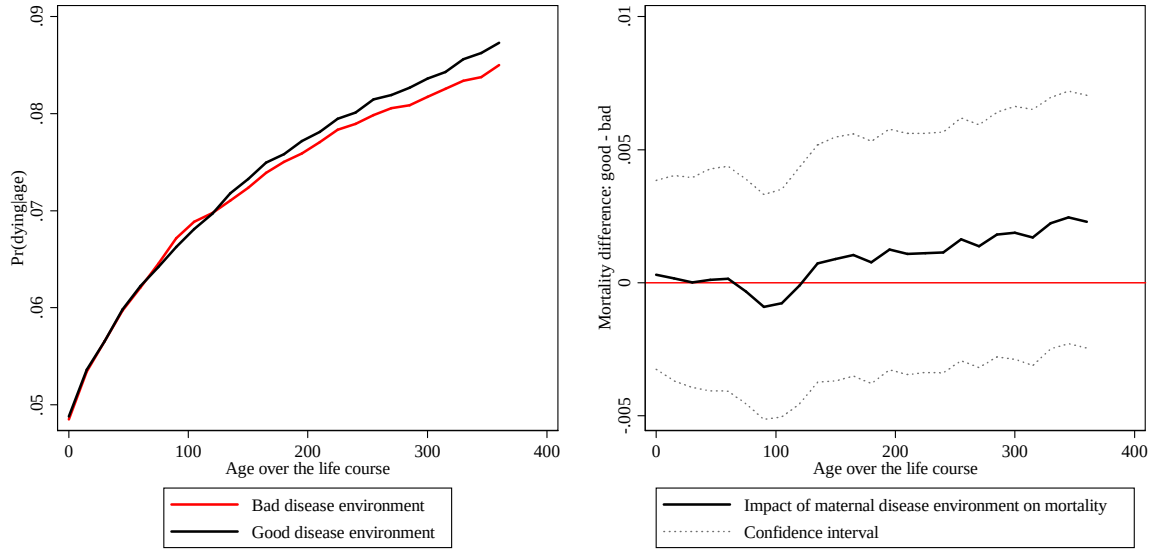
Another important issue is the population for whom the effect is identified. Even if the health shock in the first generation is random, mothers who are hit by the most severe shocks might die younger or might not get an offspring. We can only identify the effect for those first-generation individuals who survived and become fertile. Due to selection occurring among the first generation our estimate is likely a lower bound. In Section 1.5.4 we elaborate further on the transmission channels in order to be able to better assess the quality of the regression results.

1.5 Results

This section presents the results from the empirical analysis. We first provide some descriptive and graphical evidence for the main explanatory variable and the main outcomes. We then turn to regression estimates for the second generation. Finally, we deal with issues related to selection and confounding factors, and investigate the impact of the original health shock within the first generation.

1.5.1 Explorative graphical analysis

We now provide some visual evidence on the relationship between the maternal disease environment and survival prospects in the second generation. Since infant survival has been shown to be affected by intergenerational transmission (Almond et al., 2012), we start by zooming in on the first 365 days of life. Figure 1.4 provides two hazard plots. The left-hand side figure shows the cumulative hazard (by 15-day bins) for children born to mothers from different health environments (defined in this case as a dummy variable indicating positive and negative deviations from the local trend), whereas the right-hand side shows the difference in the cumulative hazards between the two environments.



Note: The left panel depicts the cumulated hazard rate over the life course (bin: 15 days). The right panel depicts the difference of those cumulated hazard rates between good and bad maternal disease environment at birth.
Interval computed at the 90% confidence level. Standard errors computed based on 100 bootstrap replications.

Figure 1.4: Second generation mortality hazards in the first year of life.

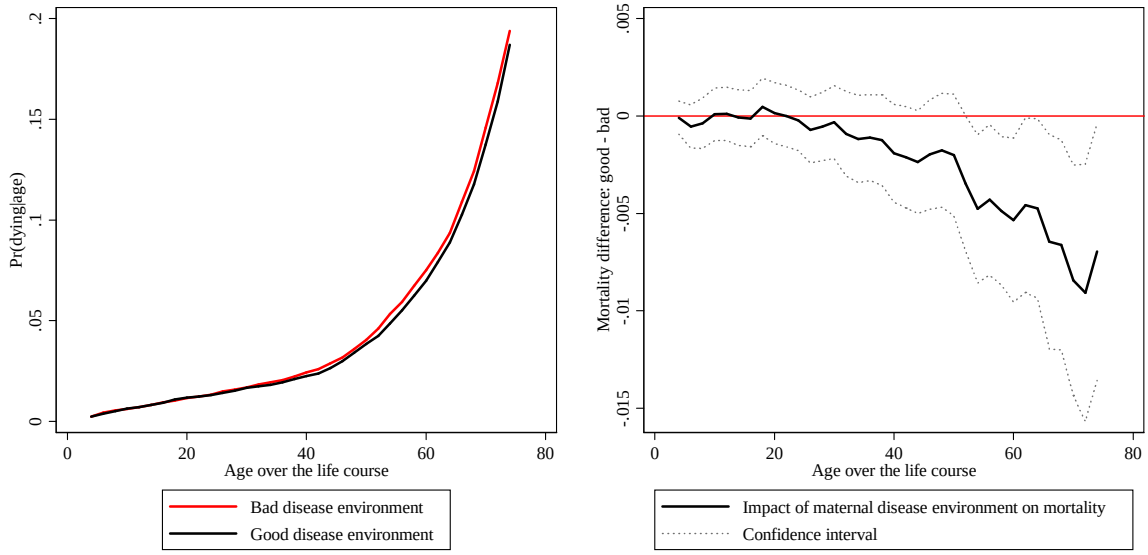
The figure does not deliver any evidence supporting the hypothesis that children born to disadvantaged mothers fare worse in terms of early life survival. We can bound the overall effect on second-generation IMR at less than 0.5 percentage points in any direction, and the point estimate is everywhere very close to zero.

Figure 1.5 shows the same exercise for the entire life course. One again the evidence suggests that the impact of the maternal disease environment is close to zero for a long time – but from age 50 onwards there is evidence of an increasing penalty for children born to mothers from a poor disease environment.

1.5.2 Second generation mortality

The evidence from the above subsection suggests that there is no manifestation of an impact of a disadvantageous maternal disease environment during the first years of life, whereas there appears to be a growing disadvantage starting around the age of 50. This evidence is consistent with the Barker hypothesis, which suggests that an adverse fetal programming may cause various health problems from middle age onwards. We now formally test this in a regression framework while controlling for various environmental factors which could possibly have confounded the relationship between the maternal disease environment and second-generation outcomes. Table 1.3 presents the results.

Each specification presents results for second-generation mortality during a particular period in life, conditional on survival up to that period. In addition to the main explanatory variable – the mother's disease environment (IMR^j) – we also control for the child's sex, the degree of crisis in the birth parish in the birth year, and a set of fixed effects for the birth parish, the mother's birth region, and the mother's birth



Note: The left panel depicts the cumulated hazard rate over the life course (bin: two years). The right panel depicts the difference of those cumulated hazard rates between good and bad maternal disease environment at birth.
Interval computed at the 90% confidence level. Standard errors computed based on 100 bootstrap replications.

Figure 1.5: Second generation mortality hazards over the entire life cycle.

Table 1.3: Regression results: second generation mortality.

Dependent Variable: Mortality between Ages				
	0-1	1-50	50-70	70-
IMR^j	0.043	−0.012	0.141**	−0.045
	(0.046)	(0.037)	(0.062)	(0.077)
crisis^i	−0.009	0.015	−0.036*	−0.067**
	(0.015)	(0.013)	(0.019)	(0.028)
Female	−0.023***	−0.038***	−0.082***	−0.110***
	(0.004)	(0.003)	(0.005)	(0.007)
j's county of birth FE	✓	✓	✓	✓
j's year of birth FE	✓	✓	✓	✓
i's parish of birth FE	✓	✓	✓	✓
Baseline (%)	8.3	7.7	16.4	30.3
R-squared	0.017	0.019	0.023	0.029
N	25,010	22,940	21,081	17,632

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses clustered at the mother j year of birth-county of birth level. Dep var: mortality in different phases of life (0-1; 1-50; 50-70; 70- y.o.) for the second generation.

year. Clearly, only the mortality rate between 50 and 70 is significantly correlated with the mother's disease environment. According to this estimate, each point increase in the IMR of the maternal birth parish associates with an 0.14-point increase in the mortality rate. Since the standard deviation of the maternal IMR is 0.05, a one standard deviation increase in this variable is associated with an increase in the second generation mortality by 0.7 percentage points. With a baseline risk is 16.4 per cent, this effect is relatively large.

We take two main messages out of Table 1.3. First, there is limited evidence of selective mortality before age 50: the relatively small and insignificant effects in the first

two columns cancel out to some extent. Second, there is a strong impact of the maternal disease environment on mortality between 50 and 70. With a p value of 0.023, this result would remain statistically significant at conventional levels also after a correction for multiple testing (p value of 0.092 using the Bonferroni method).

The existing literature suggests that males are more sensitive to health insults than females, and that the socioeconomic environment into which the child is born can act as a buffer to mitigate some health shocks. In Table 1.4, we allow for a wide range of robustness checks and analyses by subpopulation.

Each column of Table 1.4 adds some additional control variables in order to assess the robustness. The rows denoted A-I represent different subsamples, defined by sex and the socioeconomic status of the head of the household in the first generation.[22] Clearly, the survival disadvantage observed in Table 1.3 is robust to the inclusion of region-specific trends in the first generation, parental occupation, and birth order effects. The parameter drops slightly but remains statistically significant at the 5 per cent level.

We also find evidence of effect hereogeneity. When splitting the sample by sex, it becomes clear that the negative effect is entirely driven by males. When splitting by parental SES, the effect seems to be concentrated in higher socioeconomic groups. When we interact the two dimensions, the effect is particularly pronounced amongst males born in families of relatively high SES. An increase in the first-generation IMR by one standard deviation increases mortality between 50 and 70 by almost two percentage points (1.91) in this group. Measured against a baseline rate of 18.5 per cent, this effect is indeed sizeable.[23]

What are the mechanisms responsible for these findings; in particular the inverted SES gradient amongst males? We start by considering specific death causes, and then turn to socioeconomic outcomes. Since we find that the original health insult appears to interact with the early life environment of the second generation – as captured by the household head SES – it seems reasonable to consider leading death causes which are likely to have a genetic component while at the same time being modifiable. For instance, data from the death cause register enable us to differentiate between cardiovascular diseases with concomitant risk factors[24] and other cardiovascular diseases with no modifiable risk factors.

Several things are immediately clear from the results in Table 1.5. First, cardiovascular (CDV) diseases without concomitant risk factors (Panel B) appear to be responsible for a big share of the intergenerational transmission. This death cause, responsible for 23.7 per cent of the total deaths between 50 and 70, explains about 50 per cent of

[22] We will return to the potential endogeneity of this variable in section 1.5.4.

[23] In the interest of clarity, we present effect heterogeneity results as split-sample regressions. The effect heterogeneity results we present are statistically significant. Results are also robust to the inclusion of fixed effects for the mother's birth parish: due to the large overlap with the second-generation birth parish, we left this variable out of the main specification.

[24] We have information about hypertension, diabetes and alcohol consumption. Conventional risk factors for cardiovascular diseases that can be modified and/or treated are: having a diabetes condition, hypertension condition, unhealthy dietary habits, alcohol consumption, smoking and physical inactivity (see e.g. Khot et al., 2003, for an overview). Unmodifiable risk factors include for instance age and family history.

the transmission, and the concentration amongst males and in the group with higher SES is very similar as for all-cause mortality. In addition, an inverted SES gradient concentrated amongst males is supported by the estimates for the category 'diseases of the endocrine and digestive system' which includes diabetes as the primary cause of death.

Table 1.4: Regression results: second generation mortality (50-70). Robustness and effect heterogeneity.

		Dependent Variable: Mortality 50-70					
		(1)	(2)	(3)	(4)	(5)	(6)
A. All	IMR^j	0.141**	0.136**	0.136**	0.136**	0.135**	0.133**
		(0.062)	(0.062)	(0.062)	(0.062)	(0.062)	(0.062)
	R-squared	0.023	0.026	0.027	0.028	0.028	0.029
	N	21,081	21,081	21,081	21,081	21,081	21,081
B. Females	IMR^j	0.022	0.016	0.018	0.018	0.017	0.012
		(0.074)	(0.074)	(0.074)	(0.074)	(0.074)	(0.074)
	R-squared	0.019	0.027	0.028	0.028	0.029	0.032
	N	10,561	10,561	10,561	10,561	10,561	10,561
C. Males	IMR^j	0.254***	0.246**	0.239**	0.243**	0.242**	0.241**
		(0.098)	(0.097)	(0.098)	(0.098)	(0.098)	(0.099)
	R-squared	0.020	0.026	0.028	0.028	0.028	0.032
	N	10,520	10,520	10,520	10,520	10,520	10,520
D. Mid-high SES	IMR^j	0.217**	0.214**	0.214**	0.215**	0.214**	0.211**
		(0.097)	(0.096)	(0.096)	(0.096)	(0.096)	(0.097)
	R-squared	0.028	0.035	0.035	0.035	0.035	0.038
	N	10,223	10,223	10,223	10,223	10,223	10,223
E. Low SES	IMR^j	0.079	0.079	0.079	0.077	0.076	0.076
		(0.077)	(0.077)	(0.077)	(0.078)	(0.078)	(0.078)
	R-squared	0.035	0.041	0.041	0.042	0.042	0.046
	N	10,858	10,858	10,858	10,858	10,858	10,858
F. Mid-high SES, females	IMR^j	0.064	0.057	0.057	0.062	0.060	0.048
		(0.115)	(0.115)	(0.115)	(0.115)	(0.115)	(0.115)
	R-squared	0.039	0.052	0.052	0.053	0.053	0.058
	N	5,111	5,111	5,111	5,111	5,111	5,111
G. Low SES, females	IMR^j	0.009	0.011	0.011	0.010	0.008	0.013
		(0.097)	(0.098)	(0.098)	(0.098)	(0.098)	(0.099)
	R-squared	0.035	0.048	0.048	0.049	0.049	0.056
	N	5,450	5,450	5,450	5,450	5,450	5,450
H. Mid-high SES, males	IMR^j	0.384***	0.379***	0.379***	0.383***	0.386***	0.382***
		(0.145)	(0.144)	(0.144)	(0.144)	(0.144)	(0.146)
	R-squared	0.034	0.045	0.045	0.046	0.046	0.053
	N	5,112	5,112	5,112	5,112	5,112	5,112
I. Low SES, males	IMR^j	0.159	0.166	0.166	0.166	0.163	0.155
		(0.124)	(0.125)	(0.125)	(0.126)	(0.126)	(0.127)
	R-squared	0.039	0.048	0.048	0.050	0.051	0.056
	N	5,408	5,408	5,408	5,408	5,408	5,408
i's ind. controls		✓	✓	✓	✓	✓	✓
j's county of birth FE		✓	✓	✓	✓	✓	✓
j's year of birth FE		✓	✓	✓	✓	✓	✓
i's parish of birth FE		✓	✓	✓	✓	✓	✓
j's county x time trends			✓	✓	✓	✓	✓
i's parent occupation FE				✓	✓	✓	✓
i's order of birth & twin dummies					✓	✓	✓
i's year/quarter of birth						✓	✓
i's mother age FE							✓

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses clustered at the mother j year of birth-region of birth level. Individual controls include: crisis indicator and female (specifications not split by sex only). Fixed effects on i's parent occupation included in specifications not split by SES only. Dep. var: mortality between 50-70 y.o. for the second generation; sample conditional on survival until 50 y.o.

Table 1.5: Regression results: mortality by cause of death. Robustness and effect heterogeneity.

		All	By Sex		By SES		By Sex and Socio-Economic Status			
			Females	Males	Mid-High	Low	Females, Mid-High SES	Females, Low SES	Males, Mid-High SES	Males, Low SES
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
A. All-cause mortality	IMR^j	0.133**	0.012	0.241**	0.211**	0.076	0.048	0.013	0.382***	0.155
		(0.062)	(0.074)	(0.099)	(0.097)	(0.078)	(0.115)	(0.099)	(0.146)	(0.127)
	Prevalence	0.164	0.122	0.205	0.150	0.176	0.116	0.128	0.185	0.224
B. Cardiovascular diseases (no preventable/ curable risk factors)	IMR^j	0.066*	0.039	0.090	0.126*	0.028	0.054	0.023	0.202*	0.038
		(0.035)	(0.032)	(0.063)	(0.066)	(0.035)	(0.054)	(0.041)	(0.119)	(0.056)
	Prevalence	0.039	0.021	0.056	0.036	0.042	0.020	0.022	0.051	0.061
C. Cardiovascular diseases (preventable/ curable risk factors)	IMR^j	−0.002	−0.024	0.016	−0.006	0.001	−0.056**	0.002	0.047	−0.001
		(0.017)	(0.023)	(0.025)	(0.021)	(0.025)	(0.025)	(0.037)	(0.035)	(0.037)
	Prevalence	0.011	0.005	0.016	0.009	0.013	0.004	0.007	0.013	0.019
D. Other causes: Diseases of the digestive/ endocrine system incl. diabetes	IMR^j	0.016	0.016	0.018	0.036**	−0.001	0.039	0.001	0.046*	−0.003
		(0.011)	(0.016)	(0.017)	(0.018)	(0.014)	(0.026)	(0.016)	(0.025)	(0.024)
	Prevalence	0.007	0.005	0.008	0.006	0.007	0.006	0.005	0.006	0.010
Lung & oral cavity cancer	IMR^j	−0.027*	−0.031*	−0.023	−0.026	−0.029	−0.024	−0.028	−0.022	−0.029
		(0.015)	(0.019)	(0.022)	(0.021)	(0.021)	(0.027)	(0.024)	(0.029)	(0.033)
	Prevalence	0.009	0.007	0.012	0.009	0.010	0.006	0.007	0.011	0.013
Cancer (excl. lung & oral cavity cancer)	IMR^j	0.016	−0.038	0.073*	0.042	−0.015	−0.014	−0.051	0.113	0.029
		(0.026)	(0.037)	(0.039)	(0.046)	(0.034)	(0.057)	(0.053)	(0.070)	(0.045)
	Prevalence	0.039	0.041	0.037	0.038	0.040	0.041	0.041	0.036	0.039
Symptoms/signs not elsewhere classified	IMR^j	0.027	0.028	0.027	−0.042**	0.082**	−0.041	0.090*	−0.048	0.085*
		(0.022)	(0.030)	(0.032)	(0.020)	(0.036)	(0.027)	(0.053)	(0.032)	(0.049)
	Prevalence	0.010	0.009	0.011	0.010	0.011	0.008	0.010	0.011	0.012
External causes/ infections/par. diseases	IMR^j	−0.000	0.019	−0.021	0.019	−0.016	0.058*	−0.018	−0.025	−0.011
		(0.021)	(0.018)	(0.037)	(0.025)	(0.032)	(0.034)	(0.017)	(0.039)	(0.057)
	Prevalence	0.013	0.008	0.019	0.012	0.014	0.008	0.008	0.016	0.021
Other circulatory system diseases (excl. B. and C.)	IMR^j	0.037	0.001	0.072	0.032	0.041	−0.015	0.016	0.086*	0.071
		(0.025)	(0.016)	(0.047)	(0.024)	(0.040)	(0.020)	(0.026)	(0.047)	(0.074)
	Prevalence	0.012	0.008	0.017	0.011	0.013	0.007	0.008	0.016	0.019
Respiratory diseases	IMR^j	−0.013	−0.024	−0.012	0.007	−0.027	0.002	−0.035	−0.001	−0.034
		(0.014)	(0.019)	(0.021)	(0.023)	(0.018)	(0.033)	(0.023)	(0.031)	(0.029)
	Prevalence	0.011	0.009	0.013	0.010	0.011	0.008	0.009	0.012	0.014
All other causes	IMR^j	0.015	0.025	0.002	0.021	0.016	0.041	0.015	−0.012	0.015
		(0.018)	(0.026)	(0.025)	(0.032)	(0.020)	(0.053)	(0.023)	(0.038)	(0.034)
	Prevalence	0.013	0.010	0.016	0.011	0.015	0.008	0.011	0.013	0.019
N		21,081	10,561	10,520	10,223	10,858	5,111	5,450	5,112	5,408

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses clustered at the mother j year of birth-county of birth level. Dep. var: death causes for the second generation (conditional on surviving until 50 y.o.). Category "All other causes" is a residual category for all death causes not included in the other reported categories. Overall effect in col. 1, split by sex (2-3), split by SES (4-5), interaction sex*SES (6-9). All regressions include i's individual controls and fixed effects for j's county of birth, j's year of birth, i's parish of birth, the occupation of i's parents, i's birth order and twin dummies, i's year and quarter of birth, and i's mother age as well as j's county specific time trends.

These findings are informative regarding the mechanism driving the results, since these conditions – cardiovascular disease and diabetes – and the related risk factors hypertension and obesity are the main outcomes of adverse fetal programming according to the Barker hypothesis (Barker, 1990). Importantly, such metabolic adaptations in the 2nd generation are compatible with intergenerational transmission (Langley Evans, 2015).

The onset of metabolic problems is likely determined early in life and several epidemological studies show that a high-nutrient diet in infancy, in particular diets rich in dairy proteins, associates with programming of the metabolic syndrome in children – specifically with increased BMI and obesity in childhood (Weber et al., 2014; Pearce and Langley Evans, 2013). A household survey conducted 1936–7 provides detailed information on dietary habits in Swedish families during the first years of life of the second generation. The survey concludes that young children from higher SES background were given on average more nutritious food than children from lower SES background. The survey also shows that milk and, more generally, dairy products represent the type of protein that is consumed the most. Also in this case there is clear socioeconomic difference among children: a five year old child in a middle- or high-SES household on average consumed 4.6 litres milk per week – 25 percent more compared to a peer from a low-SES household (Boalt, 1939).[25] Similar indications are provided by larger, but less detailed, surveys conducted in the 1930's (Socialstyrelsen, 1938; Medicinalstyrelsen, 1934). Also, while sugar never previously had been an everyday consumption good, it becomes all the more common among higher SES-housholds in Sweden during this time period (Bolin, 1934; Torell, 2013). It is thus possible to explain a substantial part of the inverted socioecomic gradient with reference to the thrifty phenotype: second generation high-SES individuals were more likely to be exposed to a calorie rich diet and a sedentary lifestyle than low-SES individuals during their childhood.

However, Table 1.5 suggests that also behavioural factors matter to the intergenerational transmission. For example cancers appear to be responsible for some of the male penalty in intergenerational transmission, and the most prevalent cancers affecting males have a strong behavioural component. Lung and oral cavity cancers appear to counteract the intergenerational transmission, and the estimated reduction in the prevalence of this death cause is large, considering that less than one per cent of the sample died between 50 and 70 due to this cause. Besides, the absence of an SES gradient in females appears to be attributable to some extent to a reduction in cardiovascular disease with concomitant risk factors within the high-SES group.

We hence conclude that the fetal programming to some extent also appears to interact with behavioural changes in adulthood. This is to be expected: the birth cohorts under study were some of the first to become aware of the perils of smoking in adult ages (the Surgeon General's report on smoking and health was published in 1964; Holford et al., 2014). Likewise, these cohorts were the first to be exposed to large-scale prevention programmes for cardiovascular disease, which are believed to have contributed to a reduction in morbidity and mortality (Weinehall et al., 1999).

[25]The survey was conducted by Kooperativa Förbundet with the aim to map dietary habits across socio-economic groups, and covered 378 housholds (1163 individuals) across the countrey. Each household filed a protocol for each meal and every food intake for every household member during seven days.

The inverse socioeconomic gradient in the intergenerational transmission may consequently be due to a combination of environmental and behavioural factors in childhood and adulthood being of greater importance for males. This line of reasoning seems to be supported by trends in mortality rates, as Figure 1.6 illustrates.

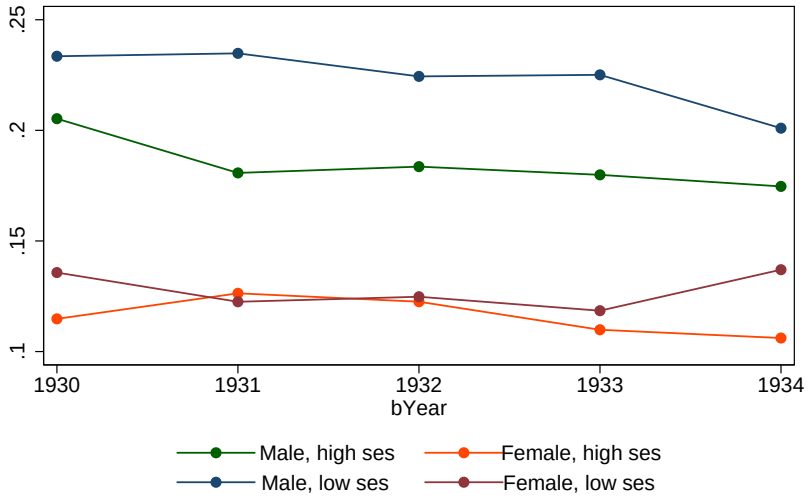


Figure 1.6: Cohort trends in mortality between 50 and 70, by sex and socioeconomic status (SES).

Figure 1.6 shows that males exhibit a strong downward trend in mortality as well as a strong socioeconomic gradient, whereas none of these features is visible for females. Our tentative conclusion is thus that the convergence of male mortality rates with female rates which this generation experienced, was to a large extent driven by improvements for higher-SES individuals who did not carry a disadvantage related to a health shock in the previous generation. This explanation is very similar in spirit to the ideas brought forward by Cutler et al. (2006), who postulate that new knowledge and treatment possibilities will generally benefit the higher SES groups first, so that the SES gradient may widen in some periods. Interestingly, the SES gradient in the intergenerational transmission of health may move in the opposite direction for exactly the same reasons.

1.5.3 Second generation SES outcomes

We now turn to regression estimates for socioeconomic outcomes of the second generation, measured at the suitable time when our individuals were between 36 and 40 years old (cf. Böhlmark and Lindquist, 2006). The SES outcomes we now consider may be seen either as mechanisms possibly explaining the findings we have reported for mortality above, or as outcomes in their own right.

Table 1.6 provides regression results for a range of specifications. Each column is equivalent to the corresponding columns in Table 1.4 and the split-sample estimates presented in the rows are consistently defined. The number of observations differs slightly due to migration (those leaving Sweden before 1970 are not observed) and mortality (individuals who died between 1970 and the age of 50 are included here but not in Table 1.4).

The statistical significance is generally somewhat weaker for earnings compared to mortality, but a clear picture nevertheless emerges: there is evidence of a relatively large penalty associated with the maternal health shock, and the result is mainly driven by females and individuals from lower socioeconomic groups. In particular females seem to suffer from the maternal health shock: the negative earnings impact for this group is more than twice as large as the overall impact. A standard deviation change in the maternal disease environment associates with a five-percent reduction in earnings in the overall population, and with a 10-per cent reduction for females. Interestingly, there is no evidence of a disadvantage for males within the high SES group. Thus, even though the maternal disease environment leads to elevated mortality later on for this group, there is no evidence that this disadvantage is manifested in earnings in middle age.

Considering labour market participation, Table 1.7 presents results for three employment variables – for the overall sample and for males and females separately. The findings for earnings are reflected in the results for employment and, in particular for the female subgroup, a poor maternal disease environment associates with a reduction in employment.

Table 1.6: Regression results: earnings in 1970.

		Dependent Variable: Log Labor Income in 1970					
		(1)	(2)	(3)	(4)	(5)	(6)
A. All	IMR^j	-0.986*	-0.914*	-0.907*	-0.921*	-0.915*	-0.920*
		(0.539)	(0.546)	(0.544)	(0.542)	(0.542)	(0.541)
	R-squared	0.208	0.211	0.213	0.213	0.214	0.215
	N	18,566	18,566	18,566	18,566	18,566	18,566
B. Females	IMR^j	-2.183**	-2.147**	-2.121**	-2.108**	-2.083**	-2.073**
		(1.036)	(1.052)	(1.050)	(1.050)	(1.049)	(1.048)
	R-squared	0.027	0.036	0.039	0.039	0.040	0.043
	N	9,104	9,104	9,104	9,104	9,104	9,104
C. Males	IMR^j	-0.002	0.095	0.108	0.081	0.088	0.082
		(0.425)	(0.420)	(0.422)	(0.423)	(0.421)	(0.422)
	R-squared	0.024	0.031	0.033	0.034	0.035	0.040
	N	9,462	9,462	9,462	9,462	9,462	9,462
D. Mid-high SES	IMR^j	-0.354	-0.409	-0.409	-0.424	-0.401	-0.481
		(0.759)	(0.763)	(0.763)	(0.760)	(0.758)	(0.757)
	R-squared	0.227	0.232	0.232	0.233	0.233	0.236
	N	10,034	10,034	10,034	10,034	10,034	10,034
E. Low SES	IMR^j	-1.480*	-1.310	-1.310	-1.359*	-1.377*	-1.333*
		(0.802)	(0.808)	(0.808)	(0.806)	(0.805)	(0.803)
	R-squared	0.204	0.209	0.209	0.210	0.211	0.214
	N	8,532	8,532	8,532	8,532	8,532	8,532
F. Mid-high SES, females	IMR^j	-1.667	-1.692	-1.692	-1.656	-1.563	-1.632
		(1.477)	(1.479)	(1.479)	(1.474)	(1.476)	(1.474)
	R-squared	0.048	0.062	0.062	0.063	0.065	0.070
	N	4,926	4,926	4,926	4,926	4,926	4,926
G. Low SES, females	IMR^j	-2.426	-2.236	-2.236	-2.258	-2.281	-2.103
		(1.492)	(1.525)	(1.525)	(1.523)	(1.522)	(1.516)
	R-squared	0.046	0.061	0.061	0.062	0.063	0.070
	N	4,178	4,178	4,178	4,178	4,178	4,178
H. Mid-high SES, males	IMR^j	0.326	0.325	0.325	0.298	0.299	0.218
		(0.487)	(0.488)	(0.488)	(0.489)	(0.490)	(0.493)
	R-squared	0.044	0.056	0.056	0.057	0.058	0.063
	N	5,108	5,108	5,108	5,108	5,108	5,108
I. Low SES, males	IMR^j	-0.390	-0.159	-0.159	-0.185	-0.186	-0.164
		(0.690)	(0.675)	(0.675)	(0.677)	(0.676)	(0.677)
	R-squared	0.044	0.059	0.059	0.060	0.061	0.073
	N	4,354	4,354	4,354	4,354	4,354	4,354
i's ind. controls		✓	✓	✓	✓	✓	✓
j's county of birth FE		✓	✓	✓	✓	✓	✓
j's year of birth FE		✓	✓	✓	✓	✓	✓
i's parish of birth FE		✓	✓	✓	✓	✓	✓
j's county x time trends			✓	✓	✓	✓	✓
i's parent occupation FE				✓	✓	✓	✓
i's order of birth & twin dummies					✓	✓	✓
i's year/quarter of birth						✓	✓
i's mother age FE							✓

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses clustered at the mother j year of birth-region of birth level. Individual controls include: crisis indicator and female (specifications not split by sex only). Fixed effects on i's parent occupation included in specifications not split by SES only. Dep. var: log labor income measured in 1970 (from Census 1970).

Table 1.7: Regression results: employment in 1970.

	All			Females			Males		
	Any (1)	Fulltime (2)	Part-time (3)	Any (4)	Fulltime (5)	Part-time (6)	Any (7)	Fulltime (8)	Part-time (9)
IMR^j	-0.110*	-0.008	-0.102**	-0.355***	-0.142	-0.213**	0.080	0.086	-0.006
	(0.061)	(0.061)	(0.047)	(0.111)	(0.109)	(0.093)	(0.054)	(0.059)	(0.028)
$crisis^i$	-0.015	-0.006	-0.009	0.018	0.038	-0.020	-0.033	-0.035	0.003
	(0.025)	(0.025)	(0.020)	(0.044)	(0.042)	(0.039)	(0.022)	(0.025)	(0.012)
Female	-0.332***	-0.571***	0.240***						
	(0.006)	(0.006)	(0.005)						
i's ind. controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
j's county of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
j's year of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
j's county x time trends	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's parent occupation FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's order of birth & twin dummies	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's year/quarter of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's parish of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's mother age FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Baseline (%)	76.7	63.3	13.4	59.5	33.8	25.7	93.3	91.6	1.6
R-squared	0.185	0.381	0.141	0.084	0.110	0.040	0.069	0.064	0.023
N	18,566	18,566	18,566	9,104	9,104	9,104	9,462	9,462	9,462

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses clustered at the mother j year of birth-region of birth level. Dep. var: probability of being employed in any job, in a full time job, in a part time job, respectively. Split by sex: columns 4-6: females; columns 7-9: males.

The results for 1970 earnings thus produce an additional intriguing result – that the economic disadvantage of a poor maternal disease environment is disproportionately suffered by women – but no support for the previous finding for high-SES males. In an attempt to resolve the issue we turn to education. Table 1.8 reports estimates with years of schooling in 1970 as the outcome variable. Estimates are not statistically significant at conventional levels and do not show a consistent pattern when looking at effect heterogeneity. In contrast to Richter and Robling (2013) we do not note any evidence of an education effect of the maternal health insult and we again find evidence suggesting that most of the effects of this kind of health insult manifest themselves only in adulthood.

Since we can rule out that the maternal health insult had a large impact on choices regarding education early in life, the strong earnings disadvantage of low-SES females must have other causes. The cohorts we consider lived through the expansion of the welfare state and the associated improvements in labour market opportunities for females (Magnusson, 2010). It thus seems natural to hypothesise that this trend expanded the opportunities disproportionately for females, and that the transition into formal employment within this group was related to health. In Table 1.9, we formally test this hypothesis by regressing the maternal disease environment on employment in public services. Estimation results are presented in Table 1.9.

Table 1.8: Regression results: years of schooling.

	All	**By Sex**		**By Socio-Economic Status**		**By Sex and Socio-Economic Status**			
		Females	Males	Mid-High	Low	Females, Mid-High SES	Females, Low SES	Males, Mid-High SES	Males, Low SES
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
IMR^j	0.258	0.460	0.039	0.545	0.074	1.044	-0.023	-0.002	0.281
	(0.382)	(0.511)	(0.542)	(0.666)	(0.487)	(0.950)	(0.634)	(0.875)	(0.712)
crisis^i	-0.000	0.219	-0.179	0.007	-0.039	0.193	0.056	-0.048	-0.137
	(0.137)	(0.200)	(0.208)	(0.214)	(0.186)	(0.325)	(0.258)	(0.307)	(0.286)
Female	-0.145***			-0.068	-0.200***				
	(0.034)			(0.051)	(0.046)				
i's ind. controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
j's county of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
j's year of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
j's county x time trends	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's parent occupation FE	✓	✓	✓						
i's order of birth & twin dummies	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's year/quarter of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's parish of birth	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's mother age FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Baseline	9.588	9.518	9.657	9.698	9.485	9.664	9.382	9.731	9.587
R-squared	0.100	0.117	0.115	0.116	0.060	0.138	0.094	0.148	0.081
N	20,518	10,139	10,379	9,919	10,599	4,903	5,236	5,016	5,363

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses clustered at the mother j year of birth-region of birth level. Dep. var: years of education in 1970 for the second generation.

Table 1.9: Regression results: employment in public services in 1970.

	All	**By Sex**		**By Socio-Economic Status**		**By Sex x Socio-Economic Status**			
		Females	Males	Mid-High	Low	Females, Mid-High SES	Females, Low SES	Males, Mid-High SES	Males, Low SES
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
IMR^j	-0.076	-0.158*	-0.009	-0.026	-0.105	-0.046	-0.252**	0.027	0.015
	(0.050)	(0.083)	(0.063)	(0.080)	(0.067)	(0.132)	(0.110)	(0.094)	(0.088)
crisis^i	0.000	0.019	-0.013	-0.007	-0.006	0.004	0.005	-0.013	-0.008
	(0.020)	(0.034)	(0.025)	(0.030)	(0.030)	(0.052)	(0.048)	(0.034)	(0.040)
Female	0.073***			0.086***	0.062***				
	(0.005)			(0.007)	(0.007)				
i's ind. controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
j's county of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
j's year of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
j's county x time trends	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's parent occupation FE	✓	✓	✓						
i's order of birth & twin dummies	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's year/quarter of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's parish of birth	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's mother age FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Baseline (%)	12.3	16	8.6	13.2	11.2	17.4	14.5	9.1	8.1
R-squared	0.047	0.050	0.052	0.056	0.050	0.064	0.080	0.075	0.062
N	18,566	9,104	9,462	10,034	8,532	4,926	4,178	5,108	4,354

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses clustered at the mother j year of birth-region of birth level. Dep. Var: being employed in public services in 1970, overall effect (1), split by sex (2-3), split by SES (4-5) and interactions sex*SES (6-9).

The results do indeed support a story according to which the emerging welfare state selectively employed females (in particular low-SES females) with a positive health transmission from the previous generation. We may thus conclude that the female penalty is driven by changes on the labour market, whereas the male penalty is largely unrelated to education and labour market performance.

1.5.4 Tracing the health insult

Since the variation exploited for the identification of an intergenerational transmission of health is non-experimental, it is important to consider omitted variables and other confounding factors. Besides, it is important to understand where the variation in the in utero maternal health environment comes from. Selective mortality in the second generation was addressed in section 1.5.1. However, we conditioned on variables which are determined only in adulthood for the first generation, and which may thus well be affected by the original health insult. This would lead to the familiar bad control problem (Angrist and Pischke, 2008) and thus it is important to investigate this issue further.

In a first step, we analyse whether relevant second-generation observables at birth vary systematically with the maternal disease environment. Table 1.10 presents estimates that are all very small and generally insignificant. Consequently there seems to be no effect on first-generation SES, but given that we find strong effects on second-generation SES (Tables 1.6 and 1.7) we want to further pursue this issue. Reliable information on the grandparental SES is not available, but the mothers' maiden names are contained in the data, and these do to some extent signal social position (Clark, 2012). Similar to Clark, we extract surnames signalling higher classes – noble names; latinised surnames; typical bourgeois names – to construct a very crude measure of the SES of the maternal grandparents. The bulk of noble surnames dates back to the 17th and 18th centuries, whereas the latinised surnames are partly from the pre-industrial period and partly from the decades around 1900. Bourgeois surnames typically date from the 19th century. Our categorisation thus likely captures a combination of the grandparental SES and the SES of earlier ancestors. Given the low degree of social mobility before the industrial era (Lundh, 1999; Clark, 2012), this might be an acceptable proxy for the social origin of the mother.

Figure 1.7 plots histograms of the maternal disease environment by grandparental SES. The health insult suffered by the mother is symmetric and almost equally distributed for those coming from low- and high-SES families, respectively.

Thus, the health shock appears to be unrelated to SES. Given rigid social structures in the parental generation, one might conjecture that the same holds for the parental SES, even though it is determined after the health insult. We investigate this hypothesis in Figure 1.8 by contrasting socially mobile individuals to the rest. Again, we do not find any evidence that the maternal disease environment is related to SES in either of the two generations.[26]

[26] We also reran all regressions above using grandparental SES instead of parental SES and even though some precision is lost due to the lower informativeness of the grandparental SES indicator, all results were qualitatively the same.

Table 1.10: Assessing the potential selection effects: selection into fertility and in utero.

Dependent var:	crisis^i	Female	Twin	Wedlock	Maternal age (yrs)	SES (low=1)
	(1)	(2)	(3)	(4)	(5)	(6)
IMR^j	-0.011	0.034	0.010	-0.063	-0.248	0.076
	(0.016)	(0.071)	(0.028)	(0.051)	(0.242)	(0.079)
j's county of birth FE	✓	✓	✓	✓	✓	✓
j's year of birth FE	✓	✓	✓	✓	✓	✓
j's county x year of birth FE	✓	✓	✓	✓	✓	✓
i's parent occupation FE	✓	✓	✓	✓	✓	
i's order of birth FE	✓	✓	✓	✓	✓	✓
i's year/quarter of birth FE	✓	✓	✓	✓	✓	✓
i's parish of birth	✓	✓	✓	✓	✓	✓
i's mother age FE	✓	✓	✓	✓		✓
Baseline	0.05	0.48	0.03	0.89	28.8	0.5
R-squared	0.650	0.012	0.051	0.287	0.970	0.167
N	25,009	25,009	25,009	25,004	25,009	25,009

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses clustered at the mother j year of birth-county of birth level. Dep. var: various characteristics at birth of the second generation. The regression results are indicative of whether the IMR variable relates to fertility decisions.

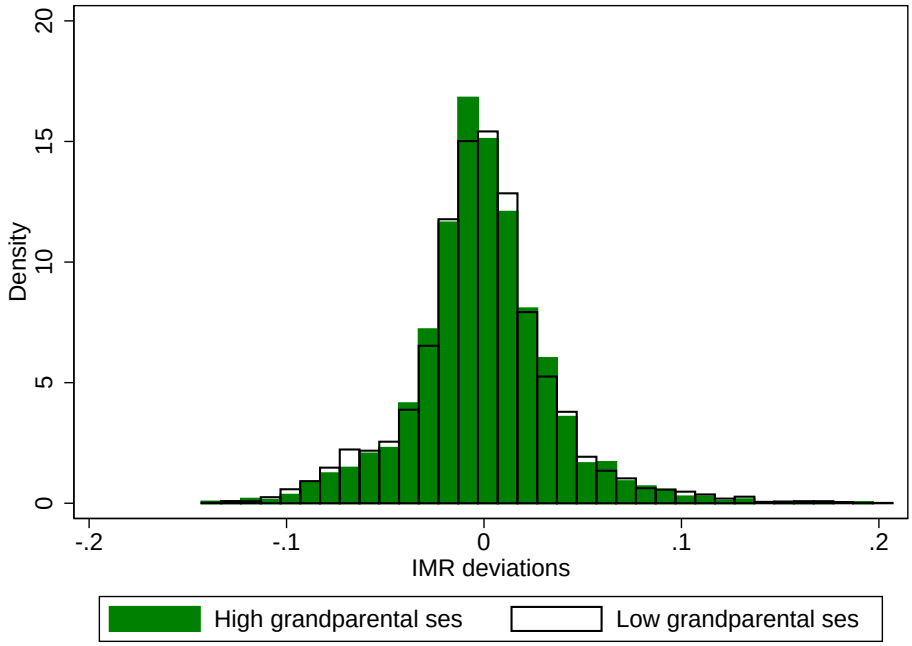


Figure 1.7: Distribution of the maternal disease environment by grandparental socioeconomic status (SES).

Next, we investigate the nature of the health shock. Relying on data from the 19th and early 20th century is a clear limitation since the available information on local public health conditions is very limited. This may however also be seen as an advantage in the sense that the local variation over time was much more random in those days, when health care services were typically neither available nor effective, and a

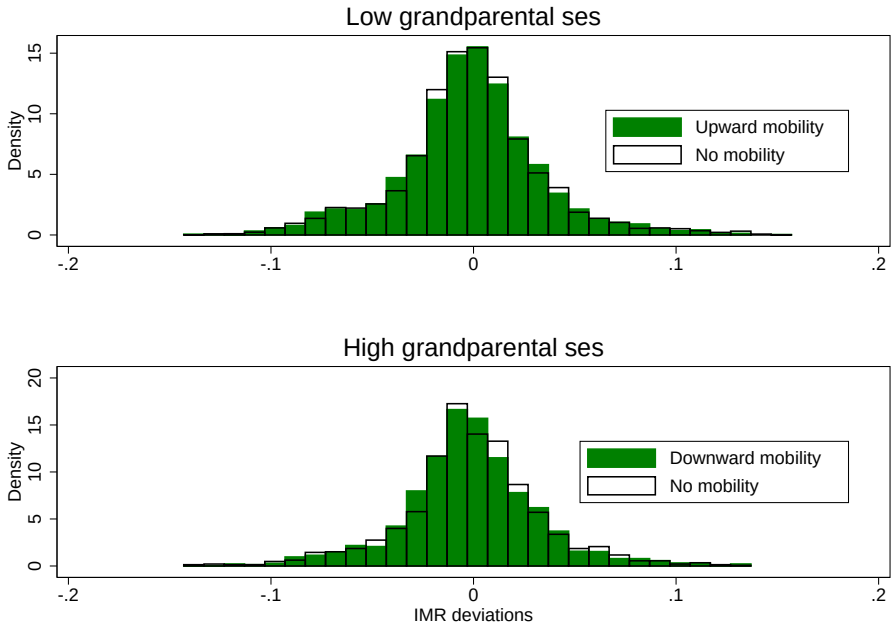


Figure 1.8: Distribution of the maternal disease environment by parental and grandparental socioeconomic status (SES).

clear socioeconomic gradient in health had not yet emerged (Bengtsson and Dribe, 2011). We now conduct an analysis at the regional level ($N = 25$) for the time period 1890-1910 and regress the regional infant mortality rate on various potential determinants, including infectious diseases and the proportion of children vaccinated against smallpox.[27]

[27]The last outbreak of smallpox in Sweden occurred in 1873–4, i.e. some years before the oldest mothers were born (Sköld, 1996). Thus, the vaccination variable serves as a proxy for the local health infrastructure and not for the disease environment.

Table 1.11: Regression of the infant mortality rate (IMR) on potential drivers (on regional level).

Dependent variable: Ln(IMR)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. Log diphtheria morbidity rate	0.0186***								0.0158**
	(0.0068)								(0.0068)
2. Log scarlet fever morbidity rate		0.0140							0.0143*
		(0.0085)							(0.0084)
3. Log respiratory disease morbidity rate			0.0093**						0.0070*
			(0.0041)						(0.0041)
4. Log share of vaccinations per capita				-0.0049					-0.0049
				(0.0032)					(0.0032)
5. Log farmhand wage					0.0030				0.0019
					(0.0059)				(0.0059)
6. Log number of pharmacies						200794*			159027
						(102522)			(103237)
7. Log midwives per females							-0.0023		-0.0006
							(0.0085)		(0.0084)
8. Log number of doctors								-0.0181*	-0.0054
								(0.0099)	(0.0037)
F-test of joint influence of disease environment (coefficients 1-3), p value in parentheses								3.99	(0.008)
F-test of joint influence of health care accessibility and living standards (coefficients 4-8), p value in parentheses								1.50	(0.188)
Region FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Year FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Region specific trends	✓	✓	✓	✓	✓	✓	✓	✓	✓
N	550	550	550	550	550	550	550	550	550

* p<0.1, ** p<0.05, *** p<0.01. Standard errors in parentheses. Note: the infant mortality rate (IMR) is measured on the regional level. Data ranges from 1889 to 1910 (22 years, 86% of the first generation mothers are born in this period) and comprises all 25 regions of Sweden.

The estimates in Table 1.11 suggest that infectious disease is the main driver of variation in regional IMRs, whereas public health resources appear to be of less importance. We also control for a proxy of living standards using the log of farm workers' wages (cf. Lundh and Prado, 2015), but it does not seem to be a relevant driver of regional IMR.

We further try to shed some light on whether our estimates capture maternal exposure in utero or during the first year of life. So far, all regressions have used the local IMR in the year of birth of the mother – and thus possibly a combination of in utero and neonatal exposure. For mothers born after 1900 we have more exact information, which allows us to exactly time the impact of the local IMR exposure. Table 1.12 compares the overall results to the results of the subset of mothers with exact date IMR information. The results clearly indicate that a large share of the estimated IMR effect is related to in utero exposure.

Table 1.12: Assessing the the underlying channel behind the second generation mortality: infant mortality rate (IMR) in utero versus in early life.

Dependent Variable: Mortality between Ages				
	0-1	1-50	50-70	70-
IMR^j	0.043	-0.012	0.141**	-0.045
	(0.046)	(0.037)	(0.062)	(0.077)
Baseline (%)	8.3	7.7	16.4	30.3
R-squared	0.017	0.019	0.023	0.029
N	25,010	22,940	21,081	17,632
IMR^j (in-utero sample)	-0.034	-0.016	0.158*	-0.099
	(0.058)	(0.063)	(0.092)	(0.117)
Baseline (%)	7.3	7.3	16.3	30
R-squared	0.022	0.024	0.035	0.036
N	11,035	10,233	9,446	7,905
IMR in-utero	-0.016	0.009	0.151*	0.128
	(0.047)	(0.051)	(0.079)	(0.104)
Baseline (%)	7.3	7.3	16.3	30
R-squared	0.022	0.024	0.035	0.036
N	11,035	10,233	9,446	7,905
j's county of birth FE	✓	✓	✓	✓
j's year of birth FE	✓	✓	✓	✓
i's parish of birth FE	✓	✓	✓	✓

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses clustered at the mother j year of birth-county of birth level. Robustness check: effect of IMR on mortality in 4 different phases of life (0-1, 1-10, 50-70, 70- y.o. in cols 1-4). First panel: benchmark (IMR measured in the year of birth of the mother); second panel: IMR in the year of birth of the mother on the subsample for which we have also IMR in-utero information; third panel: IMR in utero.

In a last set of specifications we analyse the impact of the maternal disease environment in the first generation. Table 1.13 provides the results for our main mortality outcome (mortality between ages 50 and 70). We use the indicator for grandparental SES specified above to assess whether there is heterogeneity in the impact. The overall estimated impact is small in magnitude (a one-SD increase in IMR associates with a decrease in mortality by 0.02 percentage points) but the estimates suggest there is

some heterogeneity with respect to social class: mothers of higher SES appear to suffer an increase in mortality when exposed to an unfavourable disease environment in utero, while the opposite is true for mothers from lower SES. However, the estimates are very imprecise and thus no definite conclusion may be drawn, even though the evidence seems to suggest that the impact is weaker in the first generation than in the second generation.

Table 1.13: Regression results, first generation mortality between 50 and 70.

	All	**By SES**	
		Mid-High	Low
	(1)	(2)	(3)
IMR^j	-0.035	0.012	-0.083
	(0.066)	(0.096)	(0.085)
mother_nameSES	-0.008		
	(0.008)		
j's county of birth FE	✓	✓	✓
j's year of birth FE	✓	✓	✓
i's parish of birth	✓	✓	✓
R-squared	0.017	0.030	0.026
N	15,733	7,105	8,628

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses clustered at the mother j year of birth-region of birth level. Dep. var: mortality between 50-70 y.o. for the first generation; sample conditional on survival until 50 y.o.

Thus, there is no evidence that ages 50-70 represent a particularly critical period in the first generation. But this does not necessarily imply there is no effect in adulthood at all. In order to formally test this, we conduct survival analysis for the mortality hazard of mothers from entering the sample onwards. Table 1.14 shows the results. The findings in this part clearly show evidence of scarring dominating selection in the first generation: mothers exposed to a negative disease environment did suffer elevated mortality rates in adulthood and this effect is driven by the lower socioeconomic groups. Results from Table 1.12 and Table 1.13 suggest that a socio-economic gradient in mortality appears already in the first generation – but going in the opposite direction compared to the second generation.

1.6 Conclusion

The issue of persistence of disadvantage within families has attracted great interest in economic research (Mazumder, 2005; Lindahl et al., 2012; Hertz et al., 2007). Still, evidence on the intergenerational effects of health remains scarce. More knowledge on this topic seems urgent, not the least since investments in maternal health potentially could have large returns that accumulate across generations.

Table 1.14: Cox proportional hazard regressions for mothers.

		Dependent Variable: 1st Generation Mortality			
		(1)	(2)	(3)	(4)
A. All	IMR^j	0.176	0.166	0.154	0.143
		(0.186)	(0.185)	(0.184)	(0.184)
	N.	16,961	16,961	16,961	16,961
B. Mid-high SES	IMR^j	-0.136	-0.154	-0.154	-0.172
		(0.258)	(0.259)	(0.259)	(0.260)
	N.	7,580	7,580	7,580	7,580
C. Low SES	IMR^j	0.465**	0.479**	0.479**	0.460**
		(0.222)	(0.221)	(0.221)	(0.220)
	N.	9,381	9,381	9,381	9,381
j's individual characteristics		✓	✓	✓	✓
j's county of birth FE		✓	✓	✓	✓
j's year of birth FE		✓	✓	✓	✓
Children's parish of birth		✓	✓	✓	✓
j's county x year of birth FE			✓	✓	✓
Household head occupation FE				✓	✓
N. of children FE					✓

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses. Household head occupation FE included just in the specifications not split by SES (panel A). Dep. var: mortality in the first generation (life duration in days). Cox regression (coefficients reported).

This paper contributes to a small literature on the intergenerational transmission of health and its gradient. Using historical data from Sweden for individuals born between 1930 and 1934 and their parents, and exploiting a natural health shock predicting the in utero health environment of the first generation – deviations from the local infant mortality rate – we examine the intergenerational transmission of in utero health on second generation health and socioeconomic outcomes.

In accordance with the fetal origins hypothesis the results suggest that health shocks hiting the first generation more than 100 years ago are still present in the second generation's health outcomes and thus still shape today's society: A one-standard deviation change in the mother's health environment causes the hazard to die between age 50 and 70 to increase by 0.7%.

While our results are mainly driven by males (in line with the existing literature), our data reveal an intriguing inverted SES gradient. Analysing death cause data, we find corroborating evidence that the effects to a large extent are driven by cardiovascular diseases and other conditions which have been linked to the in utero environment. The inverted SES gradient may thus be indicative of adverse effects from a rich diet (particularily during childhood), suffered by individuals exposed to poor conditions in utero. At the same time these data also suggest that the intergenerational transmission of health might be influenced by a behavioural component.

Examining socio-economic outcomes we do not find additional evidence of an inverted socio-economic gradient. On the contrary, evidence points to that females

and individuals from lower SES suffer an earnings disadvantage due to the intergenerational health transmission (a one-SD change in the IMR in the first generation associates with a 10-per cent reduction in earnings for females). Females from low SES backgrounds suffer from particularly unfavourable labor market outcomes: they are less likely to be employed and those who suffer a health insult from the previous generation are also less likely to be employed in the public sector. However, analysing years of education as an outcome we conclude that the detrimental outcomes in the second generation do not seem to emerge before individuals enter the labour market. The female penalty thus seems driven by the labour market, while the male penalty is largely unrelated to education and labour market participation.

Lastly we examine the impact of the health shock on the first generation: while there is no one-to-one transmission of the adverse effect on mortality between ages 50 and 70, when looking at the overall mortality distribution, there is some evidence that a worsening of the disease environment at birth increases the mortality risk for first generation mothers from a low SES family.

All in all, our investigation provides evidence that intrauterine programming is not only confined to one generation. It is inherited non-genetically from mothers to their children. In contrast to the existing literature that focuses on early life mortality outcomes, this paper shows that the effect on later life mortality might be even more relevant.

1.7 Appendix

Figure 1.9 suggest that the identifying assumption is likely to hold. For the sake of visibility the figure only shows parishes from those regions that are relatively well-covered in our sample. It displays the share of mothers with a "bad" intrauterine environment in a parish and it conveys one important issue: the variation in the share of mothers with a bad health environment is quite modest and they are not particularly concentrated in inter-parish clusters. We therefore argue that health shocks cannot be anticipated and that every mother in our sample can be affected ex ante.

Determinants of the IMR

Income: Attrition

We discuss attrition in the individual-level income data. We compare descriptive statistics in two subsamples: the first group consists of individuals who died before 1970, provided that they reached at least age 15, while individuals who survived up to 1970 (for whom we observe income information) form the second group. In fact, it is more reasonable to compare the subsample of 1970 survivors to the subsample of individuals who reached at least the working age, e.g., who survived until 15 y. o. Table 1.16 compares mean values for the main regressors included in our specifications by means of a t-test. Results show that some of the baseline characteristics are

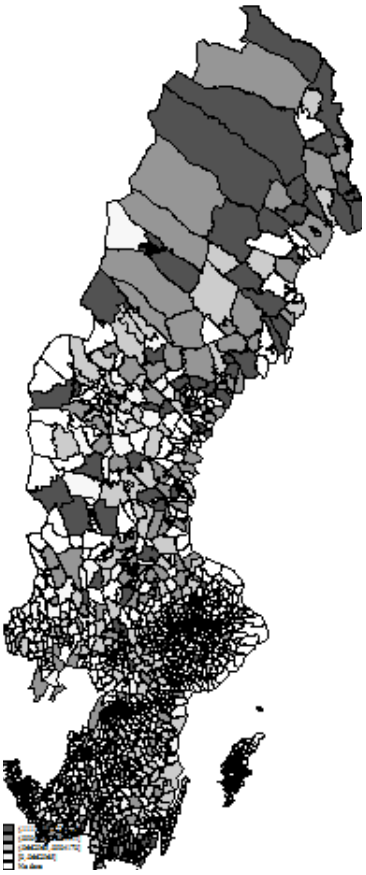


Figure 1.9: Share of mothers with a bad health environment for selected parishes.

Table 1.15: Descriptive statistics of the IMR driver analysis

Dep. var: IMR (regional 1890-1910)	mean/sd
Diphtheria morbidity rate	.0868314 (.081544)
Scarlet fever morbidity rate	.0634633 (.053987)
Respiratory disease morbidity rate	.662039 (.43812)
Salary of farmhands_	472.0387 (105.189)
Prevalence of vaccination	80.70536 (15.00039)
N	550

likely to differ between the two groups: people who died up to 1970 have, on average, a slightly higher value of the economic shock suffered during the crisis and they are more likely to have a single mother.

Table 1.16: Characteristic comparison: subsample dying before 1970 and subsample surviving to 1970.

	Died up to 1970		Alive in 1970			
Variable	N_1	μ_1	N_2	mu_2	$\mu_1 - \mu_2$	se
IMR^j	1,358	0.097	18,566	0.097	-0.000	(0.001)
crisis^i	1,358	0.062	18,567	0.046	0.016***	(0.006)
Female	1,358	0.508	18,567	0.490	0.018	(0.014)
Wedlock	1,377	-0.505	18,601	0.760	-1.265***	(0.147)
Twin	1,358	0.021	18,567	0.022	-0.002	(0.004)

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Died up to 1970: includes individuals who reached working age (15 y.o.) and died until 1970.

To check whether attrition is an issue in our data we adjust for attrition bias using inverse probability weights and we compare our main results from Table 1.6 with the corresponding specifications estimated using IPW, shown in Table 1.17. Regressions results with and without weights are very similar to each other, suggesting that attrition does not represent a major issue in our data.

Table 1.17: Attrition in Income: Regression of log labor income in 1970 with IPW.

	All	**By Sex**		**By SES**		**By Sex x Socio-Economic Status**			
		Females	Males	Mid-High	Low	Females, Mid-High SES	Females, Low SES	Males, Mid-High SES	Males, Low SES
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
IMR^j	-0.995*	-2.233**	0.092	-0.651	-1.377*	-2.029	-2.170	0.261	-0.178
	(0.548)	(1.060)	(0.429)	(0.769)	(0.808)	(1.486)	(1.525)	(0.504)	(0.680)
crisis^i	-0.113	-0.129	-0.167	0.032	-0.284	0.114	-0.443	-0.165	-0.162
	(0.208)	(0.394)	(0.161)	(0.303)	(0.291)	(0.575)	(0.550)	(0.213)	(0.240)
Female	-3.148***			-3.262***	-3.023***				
	(0.050)			(0.068)	(0.071)				
Twin	0.088	0.169	0.018	-0.061	0.433*	-0.016	0.646*	-0.038	0.102
	(0.160)	(0.288)	(0.127)	(0.225)	(0.236)	(0.424)	(0.386)	(0.166)	(0.217)
i's ind. controls									
j's county of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
j's year of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
j's county x year of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's parent occupation FE	✓	✓	✓						
i's order of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's year/quarter of birth FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's parish of birth	✓	✓	✓	✓	✓	✓	✓	✓	✓
i's mother age FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
R-squared	0.214	0.042	0.039	0.233	0.213	0.068	0.068	0.064	0.072
N	18,262	8,946	9,316	9,801	8,461	4,804	4,142	4,997	4,319

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses clusterd at the mother j year of birth-region of birth level. Individual controls include: crisis indicator; female (specifications not split by sex only), twin and maternal marital status dummies. Fixed effects on i's parent occupation included in specifications not split by SES only.

Chapter 2

Short- and medium-term effects of informal care provision on female caregivers' health[1]

2.1 Introduction

Europe's societies are getting older. Low birthrates and population ageing due to technological progress in medicine shift the age structure towards higher shares of elderly individuals. This has strong implications for labour markets and social security systems with the long-term care sector as one important part of those. The World Alzheimer Report, for instance, expects, as a result of growing numbers of people in need of long-term care, publicly funded costs of long-term care in the European Union (EU 27) to increase from 1.2% of GDP in 2007 to 2.5% in 2060 (Alzheimer's Disease International, 2013).

Already today, costs are one reason why many governments prefer informal care (care provision of close relatives and friends) over professional formal care provision. In Germany, for instance, the public long-term care insurance paid 700€ per month in 2012 for care recipients of the highest care level who are cared by family members and 1,550€ per month to the same recipient cared by professional caregivers. Germany is a country in which long-term care is still predominantly regarded the task of the family (Schulz, 2010) and informal care is more common than in comparable states like the Netherlands (Bakx et al., 2014). More than one million official care recipients (about 46% of all) are exclusively cared by family members rendering informal care the most important part of the German long-term care system.

However, provision of informal care is both mentally and physically challenging. We, therefore, analyse the question of whether there are some hidden costs – or costs often neglected in the public debate – that make informal care provision not as economic as often thought. This could be the case if informal care provision goes along

[1]This paper is written jointly with Hendrik Schmitz and is published as: Schmitz, H., and Westphal, M. (2015). Short- and Medium-term Effects of Informal Care Provision on Female Caregivers' Health. Journal of Health Economics, 42(C), 174–185. Funding of the Frtitz Thyssen Stiftung is gratefully acknowledged.

with health impairments of the caregivers. Other costs (not considered here but heavily analysed in the economic literature[2]) are forgone income for those who leave the labour force to provide care.

The economic literature on health effects of caregiving is fairly scarce.[3] To the best of our knowledge, there are only three studies on the effect of care provision on health in a narrow sense. Coe and van Houtven (2009) estimate health effects of informal caregiving in the US using seven waves of the Health and Retirement Survey (HRS). They use sibling characteristics and the death of the mother as instrumental variables that control for selection into and out of caregiving in order to identify causal effects. They find that continued caregiving leads to a significant increase in depressive symptoms for both sexes while physical health does not seem to be affected. Do et al. (2014) use data from South Korea where informal care is quite common among females caring for their parents-in-law. The data allow identifying a health effect for daughters-in-law where selection into care is taken into account by instrumenting the informal care decision with parents-in-law's health endowment. Their findings suggest that there is an increased probability of worse physical health by providing informal care. Di Novi et al. (2013) use the first two waves of SHARE to estimate the effect of caregiving on self-rated health and quality of life, measured by the CASP-12. They find positive effects of care provision on self-rated health (seen as a measure of physical health) and mixed evidence regarding quality of life (seen as a measure of mental health).

Two further papers evaluate the relationship of caregiving and caregiver drug utilisation. On the one hand, drug intake could be seen as an objective measure of poor health. On the the other hand, it sheds light on direct costs of caregiving. Van Houtven et al. (2005) assess the impact of caring on the intake of drugs using data on caregivers for US veterans. One finding is that the intensive care margin is an important factor for drug intake. Schmitz and Stroka (2013) exploit data of a large German sickness fund that enables to consider prescriptions of anti-depressants and drugs to restore physical health. Their results support Van Houtven et al. (2005), providing some evidence that caregiving increases the intake of anti-depressants in particular if coupled with having a job. Other studies look at broader welfare consequences of caring and use life satisfaction as a proxy (Bobinac et al., 2010, Van den Berg and Ferrer-i Carbonell, 2007, Leigh, 2010, Van den Berg and Pinger, 2014). One issue with these studies is that they do not address reverse causality and selection problems based on time-varying unobserved heterogeneity.

We use representative household data from the German Socio-Economic Panel to estimate the effects of informal care provision on female caregivers' health. The outcome variables are mental and physical summary scale measures (called MCS and PCS) for the years 2002 to 2010 that capture the multidimensional nature of health.

[2] E.g. Carmichael and Charles, 2003; Heitmueller, 2007; Heitmueller and Inglis, 2007; Bolin et al., 2008; Leigh, 2010; Van Houtven et al., 2013; Meng, 2013.

[3] In the medical literature, there is a fair amount of studies on the relationship of health and care provision. They mainly stem from the US (see e.g. Schulz et al., 1995; Stephen et al., 2001; Gallicchio et al., 2002; Tennstedt et al., 1992; Beach et al., 2000; Ho et al., 2009; Shaw et al., 1999; Lee et al., 2003; Dunkin and Anderson-Hanley, 1998; Colvez et al., 2002). In general, these studies use non-representative samples and widely disregard endogeneity problems. Furthermore, they often concentrate on more specific definitions of care, such as caring for people with dementia, etc.

Our contributions to the literature on health and informal care are twofold: First, we use a different approach to address selection into and out of care provision. Except for Di Novi et al. (2013), previous studies that deal with endogeneity problems all use instrumental variables approaches. We try to identify the effect of caring using different assumptions that can put the literature on a broader basis and thereby complement it. Our approach is to fully exploit the time dimension and richness of panel data in order to justify the conditional independence assumption that would allow for a causal interpretation of the results. To be more precise, we use a regression adjusted matching approach. Although we argue below that, given our data we can justify the conditional independence assumption, we allow in a sensitivity analysis that follows Ichino et al. (2008) for certain deviations from this assumption.

Second, to the best of our knowledge, this is the first study that does not only look at contemporary, or short-term effects of informal care provision on health, but also on medium-term effects of up to seven years after care provision. By medium-term effects we mean: if a women provides care in a certain year, what is her expected change in health up to seven years afterwards. This adds on work by Coe and van Houtven (2009) who also discuss persistence of health effects but need to stick to a two year period. Medium-term consequences could be more severe than instantaneous short-term health impacts restricted to the period of providing care. Moreover, knowledge about the persistence of health effects is arguably more important for policy makers than about short-run effects only.

The results suggest that there is a considerable negative short-term effect of informal care provision on mental health which, however, fades out over time. Five years after care provision the effect is still negative but smaller and insignificant. Both short- and medium-term effects on physical health are virtually zero throughout. The sensitivity analysis suggests that sensible deviations from the conditional independence assumption do not change these results.

The paper is organized as follows. Section 2.2 briefly outlines the institutional setting of long-term care in Germany. Section 2.3 discusses the empirical approach, Section 2.4 presents the data. The results are reported in Section 2.5 while Section 2.6 assesses the sensitivity of the results. Section 2.7 concludes.

2.2 Institutional background

The German social long-term care insurance system was introduced in 1995 as a pay-as-you-go system. It is financed by a mandatory pay payroll tax deduction of currently 2.35 per cent of gross labour income (2.6 per cent for employees without children). In order to qualify for benefits, individuals need to be officially defined as care recipients and be classified into one of three care levels. In care level one individuals need support in physical activities for at least 90 minutes per day and household help for several times a week. Individuals in need of more care are classified into care levels two or three, where the benefits increase in care levels.

Benefits also depend on the type of care, where monthly payments for informal care range from 235€ (level one) to 700€ (level three), for professional ambulatory care

from 450€ to 1,550€ and for professional nursing home care from 1,023€ to 1,500€ . The latter, in particular, does not fully cover the expenses for nursing home visits and copayments of up to 50 per cent are standard. Copayments for professional ambulatory care are smaller and amount to an average of 247€ or about 20 per cent (Schmidt and Schneekloth, 2011). Social welfare may step in if individuals are not able to bear the copayment. Thus, the decision for formal or informal ambulatory care is usually not driven by financial aspects as each care recipient who is assigned a care level is entitled to benefits for all kinds of care.

The introduction of the insurance system in 1995 stressed the family as the main provider of care, as it is thought to provide care cheaper, more agreeable, and more efficiently. From the care recipient's perspective, the decision to receive informal care typically expresses a preference for being cared by familiar relatives or friends. In some cases, informal care recipients are additionally supported by professional carers. These are, on average older recipients with a higher care level and, thus, a higher care burden (Schulz, 2010). Apart from the care burden, a reason for professional care can be the absence of appropriate informal caregivers, either because they chose to only participate in the labour market or because their own physical or mental health conditions prohibits the full amount of necessary care provision.

From the caregiver's perspective, affection and sense of responsibility towards a loved parent or spouse mainly drive the decision to provide care. Although the insurance benefits for informal care are often passed on to the care provider this comparably small amount cannot be regarded a financial incentive to provide care, as it is also needed to cover other expenses for care provision (see Schmidt and Schneekloth, 2011 for all points). However, the insurance funds do pay pension contributions for informal carers who provide care at least 14 hours a week (Schulz, 2010). In 2002, people cared on average 14 hours per week for care recipients whose assessment of needs is at least classified as the lowest official category (Schneekloth and Leven, 2003).

Between 2001 and 2011 there were only minor adjustments to the German long-term care system. They were minor because benefits were increased but only to keep pace with the inflation (Rothgang, 2010) and, thus, did not change the incentives to provide care. As of 2008, employed individuals are allowed to take a 10 day (not repeatable) unpaid leave to organize or provide care in case of an incidence of care dependency in the family. However, only very few caregivers make use of this.[4] Thus, the tasks of informal caregivers, the composition of caregivers and care recipients as well as financial incentives remained fairly similar over time.

2.3 Empirical strategy

We aim at estimating the effect of informal care provision on health. Certainly, the decision to provide care is not random per se. Given that someone close becomes care dependent, some individuals choose to provide care while others do not. The

[4] Schmidt and Schneekloth (2011) report that only 9,000 out of possibly 150,000 made use of this until 2011. The most frequent reason for not making use in their survey was that individuals were not aware of the possibility.

willingness to provide care depends on factors such as the financial and temporal affordability, own health endowment as well as innate tendencies such as personality traits.

To deal with this problem we apply the model of Rubin (1974). Following his notation we observe $Y = T \cdot Y_1 + (1 - T) \cdot Y_0$, where T indicates whether an individual is assigned to treatment (two hours of informal daily care provision, but we will also consider alternative definitions) or control group, Y is the outcome (health), and the index $\{0, 1\}$ indicates the potential health outcome of being a caregiver or not. If we simply compare the realized outcomes, i.e. $E(Y_1|T = 1) - E(Y_0|T = 0)$, selection bias will most likely arise due to the non-randomness of care provision. However, the average treatment effect on the treated (ATT) can be identified if the conditional independence assumption holds and assignment to treatment is random conditional on controls: $Y_1, Y_0 \perp T|X$. That is, if all the determinants that simultaneously influence the health outcome and the selection into treatment are observed. Then, $ATT = E(Y_1 - Y_0|T = 1, X)$ is the causal ceteris paribus impact of informal care provision on health.

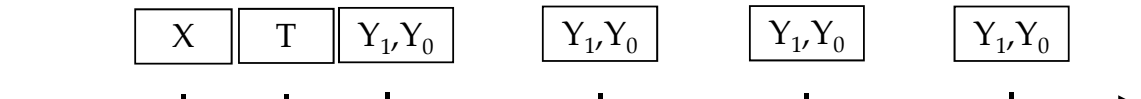
We use propensity score methods to estimate this effect and combine matching with regression methods, thus employing the so called regression adjusted matching approach (see, for example, Rubin, 1979). The advantage to using either only matching or linear regression is that it yields consistent estimates if either one of each method fails to remove the selection bias. This is called the double robustness property (Bang and Robins, 2005). Nevertheless, this method rests on the conditional independence assumption and if it does not hold and both, regression model and propensity score estimation are wrongly specified, the estimates are biased. The estimation strategy is a two-step process, originally proposed by Bang and Robins (2005). As a first step, the probability of being a caregiver (the propensity score) conditional on relevant covariates is estimated with a probit model. Subsequently, treatment and control group are matched. We use an Epanechnikov kernel with a bandwidth of 0.03 in the basic specification. To further increase the comparability, the sample is restricted to the common support of the propensity scores of the treatment and control group.

As a second step, the health outcome is regressed on informal care and, again, all control variables where the observations are weighted by the kernel weights $\boldsymbol{W}$ estimated by the matching algorithm: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}$. Standard errors are computed according to the suggestion of Marcus (2014) who employs robust standard errors of the regression above since they are slightly more conservative but easier to estimate than bootstrapped standard errors that, in addition, are not formally justified.[5] However, we cluster standard errors on the individual level since individuals appear several times in the data set.

We employ the time structure as presented in Figure 2.1. Assignment to treatment T occurs in $t = 0$. We condition on a large set of covariates in $t = -1$, thus reducing the potential problem that covariates are affected by the treatment status. We, then, compute the treatment effect four times: 1 year after treatment, 3 years after treatment, 5 years after treatment, and 7 years after treatment. Note that conditioning variables and treatment group assignment are always the same and determined

[5] We can confirm this finding in our data. Bootstrapped standard errors yield slightly less conservative standard errors.

in $t = -1$ and $t = 0$, respectively. As explained in Section 2.4, the outcome variable is available biannually between 2002 and 2010 in our data set. Since we condition on pre-treatment outcome (see explanation below), the earliest possible treatment year is 2003. We use the maximum available information in the data and pool it to one sample. Then, individuals treated in 2003 (call this wave 1) can be followed until $t = 7$ in 2010 whereas individuals treated in 2009 (call this wave 4) can only be followed until $t = 1$. Hence, the effect in $t = 1$ will be measured more precisely than the one in $t = 7$.



	t=-1	t=0	t=1	t=3	t=5	t=7
Wave 1	2002	2003	2004	2006	2008	2010
Wave 2	2004	2005	2006	2008	2010	
Wave 3	2006	2007	2008	2010		
Wave 4	2008	2009	2010			

Figure 2.1: Basic time structure

Even though we condition on a large set of covariates that are supposed to capture the process of the decision to provide care, there are probably some threats to the conditional independence assumption. First, there might be health driven selection into treatment. Individuals who are confronted with the question to provide care but are themselves in poor health might not be able to do so. As informal care provision is both physically and mentally challenging, this possible selection holds for both dimensions of health. If this is indeed the case and informal care provision has negative health effects, ignoring this reverse causality problem would lead to an underestimation of the true effects (in absolute values). We follow, e.g., Lechner (2009a) and García-Gómez (2011) and match individuals on pre-treatment outcomes (here, health status in $t = -1$), thus only comparing individuals of the same baseline health status before treatment. This rules out that individuals in the control group are in worse health due to a selection of healthy individuals into care provision.

A second issue is unobserved heterogeneity, confounders that both affect treatment and outcome, but are not observable for the researcher. As Lechner (2009a) suggests, this problem can be mitigated by stratifying the sample according to care provision in $t = -1$. Comparing only individuals with the same care status in $t = -1$ accounts for a lot of unobserved heterogeneity that affects treatment participation. Hence, the conditional independence assumption is much more likely to hold within the strata of previous care provision.[6] Moreover, stratifying the sample at least mitigates the problem that control variables, though dated back to $t = -1$, could be determined by

[6]However, for stratum 1 (individuals who already provided care in $t = -1$) there is presumably more unobserved heterogeneity left, since here, all individuals that have been caring, potentially for several years, are pooled. Thus, we identify only an average effect over all conceivable care spells. This, however, holds for all studies that cross-sectional data or panel data and do not explicitly model the dynamics of care.

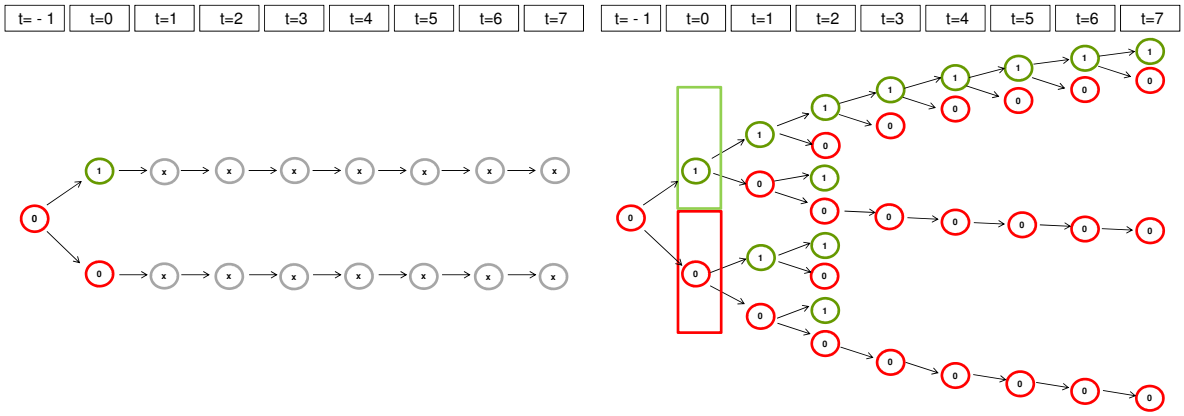
care provision in $t = 0$ through confounders that both affect past control variables and current treatment status.

Table 2.1: Stratified sample

Stratum	t=-1	t=0
1	care	care
		no care
2	no care	care
		no care

Hence, we generate two samples based on information in $t = -1$ and estimate the treatment effects independently for each sample as depicted in Table 2.1. Both estimated treatment effects and their variances for each stratum are merged as weighted means.[7]

Note that treatment is only defined as care provision in $t = 0$ while we leave future care status unrestricted as exemplarily shown in Figure 2.2a for care starters. The most important advantage of this is that selection out of care provision due to bad health is no problem in identifying medium-term effects of care provision because future health status – potentially affected by care and potentially leading to selection out of care provision in later years – does not affect the treatment group assignment at all. A drawback might be that in this static model sequential paths of care provision are not explicitly modelled. Figure 2.2b shows some examples of paths after $t = 0$. Individuals who care in $t = 0$ might either stop in $t = 1$ or go on and stop later, or even stop and take up care provision again. The same holds for the control group that includes individuals who cared later on.



(a) Modelled paths (b) Excerpt of underlying paths

Figure 2.2: Group assignment rules

Note: 1 = providing care; 0 = not providing care; X = care status not specified (= either 1 or 0). Right panel does not include all possible paths but only a small excerpt.

Lechner (2009b) and Lechner and Miquel (2010) present a dynamic matching model that enables to compare the effect of, say, providing care in each period between $t = 0$

[7] $\widehat{ATT} = \frac{1}{n}\sum_{i\in 1,2} n_i \cdot \widehat{ATT}_i$, $\widehat{se} = \sqrt{\frac{1}{n^2}\sum_{i\in 1,2} n_i^2 \cdot se_i^2}$.

and $t = 7$ with, say, not caring in any of the periods.[8] We do not make use of such a framework. Most importantly, because the numbers of observations in the 256 ($= 2^8$) different paths become very small in our sample, except for the path of never carers $(0 - 0 - \cdots - 0)$. E.g., only 23 observations in our data set provide care in each year between $t = 0$ and $t = 7$. Second, given that we condition on pre-treatment outcome, we would need to condition on the health status at each node in Figure 2.2b in order to make the "dynamic conditional independence assumption" (Lechner and Miquel, 2010) credible. However, as the outcome variable is only available in every other year, we cannot, for instance, condition on health in $t = 2$ in modelling the transition of care provision between $t = 2$ and $t = 3$. Hence, we follow, e.g., Lechner et al. (2011) and use the standard static version. This enables us to estimate the average effect of care provision in $t = 0$ on health in later years. This effect is generated by dynamics in care provision which are not explicitly modelled but implicitly taken into account. The descriptive statistics in the next section show that the vast majority of care durations is between one and three years. Hence, in reality, there is much less heterogeneity in care durations than implied by all theoretically possible paths in Figure 2.2b.

Above, we have set out selection issues and the responses to those that are facilitated by our data. To assess the adequacy of our responses we report a sensitivity analysis in Section 2.6. We estimate a short- and a medium-term effect of care provision, where by medium-term effect we mean the expected health effect of care provision in a certain year, five or seven years after. Given that an individual cannot foresee her care provision path in the future, this expected effect (though probably a composite of effects from different paths) is arguably the most relevant one from an individual perspective when deciding about providing care in $t = 0$ or not.

2.4 Data

We use data from the German Socio-Economic Panel (SOEP) which is a yearly repeated representative longitudinal survey of households and persons living in Germany that started in 1984. The SOEP covers a wide range of questions on the socio-economic status like on work, education, health, and personal attitudes (see Wagner et al., 2007, for details). Currently, some 22,000 individuals above the age of 18 from more than 10,000 households are interviewed each year.

We restrict the sample to women that have complete information on treatment status in $t = 0$ and control variables in $t = -1$. Since caregiving among men is much less common and we observe considerably fewer male caregivers we drop men, as it turned out to be very difficult to properly model the treatment participation (the propensity scores yielded only very low values). Moreover, we drop female professional caregivers from the sample, as they might mix up professional and personal affairs. Beyond that, no further restrictions are imposed on the sample. Pooling all waves as shown in Figure 2.1, we end up with a sample of 31,177 person-year observations in $t = 0$. The lowest line of Table 2.2 shows the number of observations in the

[8] See also Augurzky et al. (2012) for an application.

sample. Of the 31,177 observations in $t = 0$, we observe the health status of 28,622 in $t = 1$, of 20,288 in $t = 3$, and so on. This number strongly drops over time, mainly because more and more episodes are right censored (again, see Figure 2.1).

Table 2.2: Sample size

	t=0	t=1	t=3	t=5	t=7
Hours of care = 0	29,080	26,667	18,956	11,455	5,194
Hours of care = 1	862 (= 41%)	800	564	357	160
Hours of care = 2	507 (= 24%)	479	317	197	85
Hours of care = 3	203 (= 10%)	193	140	81	36
Hours of care = 4	167 (= 8%)	152	100	53	24
Hours of care $>$ 4	358 (= 17%)	331	211	111	53
All observations	31,177	28,622	20,288	12,254	5,552

Source: SOEP, own calculations. Number in parentheses is the share among all individuals with positive hours of care. Hours of care are measured in $t = 0$ only.

We identify caregivers depending on how individuals respond to the following question: "What is a typical day like for you? How many hours do you spend on care and support for persons in need of care on a typical weekday?" which has been included into the SOEP questionnaire since 2001.[9] Answers to this question are also shown in Table 2.2. 862 or 41% of all individuals who care a positive number of hours per day, care for one hour. 24% care two hours, whereas 10% care three hours per day. Note that the numbers in $t = 1$ (and later) do not refer to care provision in $t = 1$ but to the number of observations who care in $t = 0$ and are still observed in $t = 1$.

If an individual states caring at least two hours per day we consider her a caregiver. That is, the treatment indicator is the binary variable of caring for at least two hours per day. This comes closest to other definitions in the literature, e.g. Leigh (2010) who defines care provision as caring at least for 10 hours per week. Below we show that the results are robust to higher or lower thresholds. The question does not allow for a link between caregiver and care recipient. Hence, we have no information on the care recipient and we are not able to stratify our analysis with respect to her (e.g., in order to evaluate differences between caring for spouses or parents). This is a common shortcoming in this literature.

Table 2.3 gives a notion of the duration of care episodes. It counts the consecutive years individuals provide care of at least two hours per day. In presenting the numbers we distinguish between uncensored spells (of individuals that are observed to provide no care before and after a care episode) and censored spells (individuals that either enter the sample as caregivers or are caregivers at the end of the observation period). Due to the sample construction, there are many right censored individuals which complicates the interpretation of the table somewhat. What should be taken

[9]This question does not refer to child care which is a separate category in the time use questionnaire. The Supplementary Material includes a paragraph on the justification for the validity of self-reported answers to these kinds of questions.

away from it is that the vast majority has care spells of about one to three years. Therefore, the effects after seven years are mainly driven by individuals who had shorter caregiving episodes. Individuals who constantly care over many years hardly add to the results.[10]

Table 2.3: Care duration

Years of consecutive care as of t=0		1	2	3	4	5	6	7	8	Total
Uncensored	Observations	348	107	35	29	8	7	6	-	542
	Share	65%	19%	7%	6%	1%	1%	1%	-	100%
Censored	Observations	238	183	77	90	37	39	12	19	693
	Share	35%	26%	12%	13%	5%	5%	1%	3%	100%
thereof:										
Left censored	Observations	80	27	16	11	10	6	4	19	173
	Share	46%	16%	9%	7%	6%	3%	2%	1%	100%
Total	Observations	586	290	112	119	45	46	18	19	1,235
	Share	47%	23%	9%	10%	4%	4%	1%	2%	100%

Source: SOEP, own calculations. Uncensored individuals did not provide care in $t = -1$ and stopped caregiving some time before $t = 7$. Therefore, the maximum observable care duration is 7 years. In contrast to the empirical analysis in the rest of the paper, this table uses information up to the wave of 2011 or $t = 8$ in order to be able to calculate the number of individuals who exactly care for 7 years.

The two outcome measures are a mental and a physical health score that are based on information from the SF-12v2 questionnaire, a component of the SOEP, which includes twelve questions on mental and physical health. All items capture the general current mental and physical health status since all questions relate to the past four weeks, see the questionnaire in Table 2.6 in the Appendix. Answers to these questions are collapsed into the Mental Component Summary Scale (MCS) and the Physical Component Summary Scale (PCS) by explorative factor analysis (see, Andersen et al., 2007). Thus, both variables capture the multidimensional aspect of health. The scales range from 0 to 100, normalised to mean values of 50 and standard deviations of 10 in the 2004 reference sample. Higher values mean a better health status. MCS loads information on perceived melancholy, time pressure, mental balance and emotional problems into one summary scale.[11]

The SF-12 is commonly used to measure general health and functioning in epidemiological research (Ware et al., 1996). It includes information on subjective health but the component summary scales are correlated actual with health diagnoses. For example, Gill et al. (2007) find that MCS "is a useful screening instrument for depression

[10]This is due to the very low number of observations. Moreover, these 19 individuals caring throughout in our sample exhibit a mean MCS of 45.81 (compared to 49.38 overall). Thus, they do not affect the results in a quantitatively important way.

[11]The physical component comprises: Physical fitness (2 Questions), general health, bodily pain, role physical (2). The mental component comprises: Mental health (2), role emotional (2), social functioning, vitality. See the questionnaire in Table 2.6 in the Appendix.

and anxiety disorders in the general community, and thus, a valid measure of mental health". This view is supported by Vilagut et al. (2013) who find "acceptable results for detecting both active and recent depressive disorders in general population samples". This property could build the bridge between the short-term symptoms that are measured to longer-lasting health consequences that are thus also captured by this summary scale. Salyers et al. (2000) regard it as a valid and reliable instrument to measure health-related quality of life. Recently, MCS has also been used in the economic literature where it was shown to be correlated with, e.g., unemployment (Schmitz, 2011; Reichert and Tauchmann, 2011), and unemployment of spouses (Marcus, 2013). MCS and PCS were first introduced in the SOEP in 2002 and subsequently sampled every other year. This is why we restrict our observation period to the years 2002–2010.

We now turn to the selection of the control variables. Taking on the burden of care could theoretically be modeled as a three-stage process. Women provide care if (i) they need to. Given that they need to provide care, they (ii) must be willing to do so. Finally, (iii), they need to be able to provide care.[12]

At the first stage, the event that someone close becomes care dependent is a prerequisite of the need to provide informal care. This first stage in general depends on the age and the intra-familial social environment. We model the social environment by using indicators whether parents are alive, their age as well as the number of siblings.[13] The latter can reduce the need to provide care for frail parents as siblings could step in. Variables on this stage are sometimes employed as instruments for care provision in other studies.

At the second stage, given that someone close is in need of care, the willingness to provide care can be modeled as a function of socio-economic characteristics and personality traits. Socio-economic characteristics grouped in here are, e.g., own age, marital status, employment status, and level of education. Note, however, that family background variables might also belong to the first stage. For instance, singles do not need to care for a spouse or parents-in-law. Furthermore, we use character traits measured in the Big Five Inventory (Big5), well-known in psychology for being a proxy of human personality (see McRae and John, 1992 or Dehne and Schupp, 2007) as well as positive and negative reciprocity. Although the SOEP captures each item of the Big5 with relatively few questions in the 2005 and 2009 questionnaires, surveys revealed sufficient validity and reliability (see Dehne and Schupp, 2007). The items of the Big5 are: neuroticism, the tendency of experience negative emotions; extraversion, the tendency to be sociable; openness, the tendency of being imaginable and creative; agreeableness, the dimension of interpersonal relations and conscientiousness the dimension of being moral and organized (see Budria and Ferrer-i Carbonell, 2012). There are three questions for each of these items which are gathered on a 7-item scale. Furthermore, there is positive reciprocity, the tendency of being cooperative and negative reciprocity, the tendency of being retaliatory. For each personality measure, the score is generated by averaging over the outcome of the corresponding

[12] Note that we do not explicitly model this three-stage process but that we just have it in mind. Which variable belongs to which stage is then just a matter of interpretation.

[13] However, the number of brothers does not seem to play a role statistically. Thus, in the empirical model we only focus on the number of sisters. An alternative specification using that – among others – also uses the number of brothers can be found in the Supplementary Material.

questions per individual. Although these questions are only prompted twice in the SOEP and in years after the treatment assignment,[14] they are useful controls because these measures are supposed to be stable over a shorter period of time. The individual average of each measure is taken over all years as a proxy for time invariant personality.

Finally, on the third stage, the own health status determines the ability to provide care. As discussed in Section 2.3, we control for pre-treatment health (MCS and PCS). Moreover, we control for health satisfaction and life satisfaction. All control variables are listed in Table 2.4. Variables that might theoretically belong into the model but were not significant in the propensity score regression are left out. This holds, for instance, for income, the age of the father, the number of brothers, or calendar year dummies.

2.5 Results

2.5.1 Matching quality

Table 2.4 reports descriptive statistics of all covariates for different subgroups. It reveals that the mean as well as the standard deviation of the covariates are significantly different in the unweighted baseline sample. Column 4 gives the standardized difference between both means. Without matching almost all confounders are different at the 5% significance level between the carer and non-carer sample. In particular age, the age of the mother, and marital status exhibit large differences but also personality traits seem to be quite strong predictors of care provision. The kernel matching algorithm equalizes both samples by assigning different weights to each member of the control group. In order to compute these weights, we employ an Epanechnikov kernel with a bandwidth of 0.03. Whereas a bandwidth of 0.06 does not accomplish to equalize all covariates, a bandwidth of half the size balances every control variable to a standardized bias around 5 or less.

As regards the propensity score, the regions of common support are roughly $[0.04, 0.14]$ for the stratum of women who did not provide care in $t = -1$ and $[0.23, 0.87]$ for those who did provide care. The overlap within each stratum is good as we do not lose treatment observations by restricting the sample to the common support.[15] The low probabilities in the first stratum are simply due to the small amount of caregivers. This indicates that there is a large unobserved component determining caregiving. But we argue that this unobserved heterogeneity is not a big concern given the estimation strategy outlined in Section 2.3. Yet, there is one advantage of this fuzziness: It brings about a sufficiently large amount of observations in the control group having a similar value of the estimated propensity score. This provides a hint that the results are not sensitive to a different choice of the matching methods.

[14] The Big5 are included in the surveys in 2005 and 2009, whereas questions on negative and positive reciprocity are asked in 2005 and 2010.

[15] Of course, this also means that the required overlap condition stating that some randomness is needed is ensured in our model (see Heckman et al., 1998).

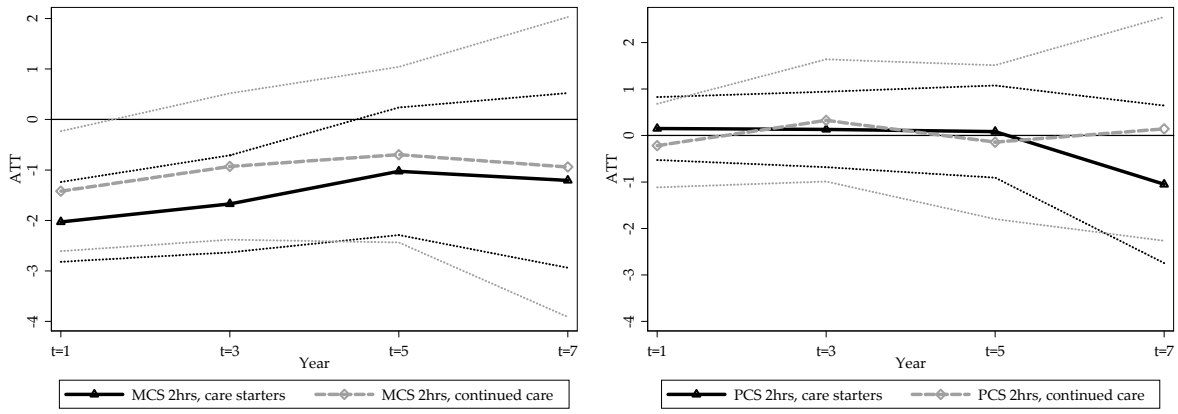
Table 2.4: Descriptive statistics according to treatment and matching status

	Treated		Controls		Matched controls		Standardized bias		
	mean	sd	mean	sd	mean	sd	unmatched sample	matched sample (0.06)	matched sample (0.03)
Stage i): care obligations									
Age of mother									
$\in [30,39]$	0.01	0.09	0.02	0.16	0.01	0.10	-13.31	-5.15	-2.35
$\in [40,49]$	0.03	0.18	0.10	0.30	0.04	0.20	-27.00	-9.33	-3.24
$\in [50,59]$	0.08	0.27	0.13	0.34	0.08	0.28	-18.15	-6.47	-2.22
$\in [60,69]$	0.12	0.32	0.12	0.32	0.11	0.32	0.25	0.96	1.12
$\in [70,79]$	0.09	0.28	0.06	0.24	0.08	0.28	9.16	4.00	1.71
$\in [80,89]$	0.09	0.28	0.02	0.14	0.08	0.28	30.58	6.81	1.15
$\in [90,99]$	0.01	0.10	0.00	0.03	0.01	0.09	12.04	3.36	2.10
Mother alive	0.46	0.50	0.48	0.50	0.46	0.50	-5.72	-3.05	-0.97
Age of father									
$\in [30,39]$	0.00	0.07	0.01	0.09	0.01	0.07	-5.14	-2.17	-1.03
$\in [40,49]$	0.02	0.12	0.08	0.27	0.03	0.16	-30.29	-10.98	-4.58
$\in [50,59]$	0.04	0.20	0.10	0.30	0.05	0.22	-20.75	-7.05	-2.12
$\in [60,69]$	0.07	0.25	0.08	0.27	0.07	0.25	-6.43	-2.13	-0.39
$\in [70,79]$	0.04	0.19	0.04	0.19	0.04	0.19	-0.40	-0.84	-0.87
$\in [80,89]$	0.01	0.11	0.01	0.10	0.01	0.11	2.34	0.44	-0.17
$\in [90,99]$	0.00	0.05	0.00	0.01	0.00	0.03	7.10	4.80	4.80
Father alive	0.19	0.40	0.34	0.47	0.21	0.41	-32.50	-11.54	-4.01
Number of sisters	1.08	1.21	1.09	1.21	1.09	1.23	-1.12	-1.28	-1.20
Partner existent	0.81	0.39	0.68	0.47	0.80	0.40	29.91	9.71	2.28
Age of partner	47.73	26.02	35.53	27.08	46.56	25.96	45.93	15.60	4.40
Stage ii): willingness to provide care									
NEURO	4.53	0.67	4.37	0.72	4.51	0.71	22.53	7.58	2.05
CONSC	6.04	0.74	5.97	0.79	6.04	0.77	9.76	2.74	0.29
AGREE	5.61	0.83	5.58	0.84	5.60	0.84	3.66	1.40	0.90
OPENN	4.37	1.15	4.51	1.12	4.40	1.13	-11.92	-4.86	-2.20
EXTRA	5.02	0.91	5.04	0.95	5.02	0.94	-1.97	-1.07	-0.49
Positive reciprocity	5.66	0.95	5.55	0.99	5.67	0.96	11.40	2.92	-0.39
Negative reciprocity	2.71	1.19	2.87	1.24	2.73	1.22	-12.81	-4.20	-1.34
Acceptance of private funding	3.31	0.81	3.29	0.8	3.31	0.82	2.68	1.12	0.44
Age	56.28	12.85	49.57	16.34	55.31	13.56	45.68	16.86	6.65
Age squared	3333.01	1419.70	2724.16	1691.40	3242.65	1471.10	38.99	14.54	5.79
Married	0.80	0.40	0.63	0.48	0.79	0.41	39.14	13.06	3.67
Divorced	0.07	0.25	0.09	0.28	0.07	0.25	-7.60	-2.67	-0.77
Single	0.07	0.25	0.17	0.38	0.08	0.27	-32.80	-11.03	-3.57
Children_hh	0.18	0.38	0.30	0.46	0.19	0.40	-29.50	-11.17	-4.27
Educ general	0.17	0.37	0.17	0.38	0.17	0.37	-2.31	-0.99	-0.73
Educ middle	0.55	0.50	0.49	0.50	0.54	0.50	11.66	4.11	1.45
Foreign	0.04	0.20	0.06	0.24	0.05	0.21	-9.91	-4.15	-2.07
West	0.69	0.46	0.75	0.43	0.70	0.46	-13.54	-5.43	-2.10
Full time	0.13	0.34	0.26	0.44	0.15	0.35	-33.85	-11.66	-3.85
Stage iii): ability to provide care									
MCS	47.38	10.52	49.47	10.12	47.59	10.91	-20.23	-6.82	-2.01
PCS	46.44	10.01	49.02	10.14	46.79	10.47	-25.57	-9.63	-3.48
Satisfaction health	6.19	2.21	6.58	2.17	6.25	2.24	-17.62	-6.56	-2.42
Satisfaction life	6.60	1.85	6.97	1.76	6.62	1.95	-20.64	-6.56	-1.07
N		1,235		29,942		29,942			

The standardized difference is calculated according to: $Diff = 100 \cdot \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{\frac{1}{2}(\sigma_1^2 + \sigma_0^2)}}$ where 0.06 and 0.03 refer to the employed Kernel bandwidth. While the bandwidth of 0.06 is only shown for sake of illustration, 0.03 is used in the estimations.

2.5.2 Estimation results

The baseline estimation results are reported in Figure 2.3 for both outcome variables MCS (2.3a) and PCS (2.3b). For convenience, we restrict this section to a graphical presentation of the results. Table 2.5 in the Appendix gives an overview of all results shown in this section. The dotted lines denote 95% confidence bands for the corresponding effect. Figure 2.3 reports the results for both pre-treatment strata separately. Care starters (black points) are those who did not care in $t = -1$ and care continuers (light grey points) those who did care in $t = -1$. The confidence bands are wider for care continuers, since this is a much smaller group. The weighted average over both effects has confidence bands comparable to the black ones in Figure 2.3.



(a) MCS (b) PCS

Figure 2.3: Baseline results MCS and PCS

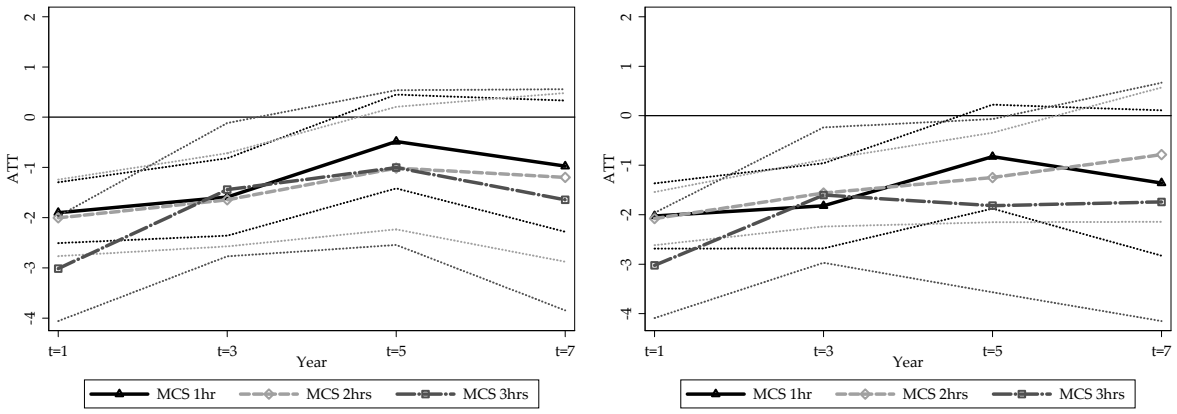
Source: SOEP. Own calculations. Note: The dotted lines indicate 95% confidence bands.

The effects are remarkably similar for both groups. If a woman cares at least two hours per day, her mental health score decreases by 2.00 units (or 20 percent of a standard deviation, sd)[16] in the first year, all other things equal. Three years after treatment assignment, this effect reduces to 16 percent of a sd before settling at below 12 per cent five and seven years afterwards. That is, women who provide care in $t = 0$ can expect to have a reduced mental health score by 12 per cent of a sd seven years after. The confidence bands indicate significant results at the 5 percent level one and three years after assignment to treatment. The effects five and seven years after are insignificant because the point estimates attenuate but in the first place because the numbers of observations strongly drop. The magnitude of the effect after seven years, however, is still 60 percent the amount of the baseline effect and thus, not neglibigle. All in all it is fair to note that, independent of the previous care status, there is a considerable short-term effect of care provision on mental health (in line with findings from previous studies, e.g. Coe and van Houtven, 2009) which decreases over time without being completely irrelevant in its extent to those who care.

In contrast, for PCS (right panel), there is basically a zero effect throughout all periods and for both strata, providing evidence for negligible effects of informal care provision on physical health. Given the absence of physical health effects, we restrict our

[16] For convenience we already report the average effect over both groups here.

analysis to mental health in the following. Moreover, we only report averaged effects of both strata of care provision in $t = -1$. Figure 2.4 presents the results for alternative daily care intensities and different definitions of the control group.[17] In Figure 2.4a we compare the effect when care provision of at least two hours per day are used to define the treatment indicator (light grey-dashed line, the baseline specification) with one hour per day (black line) and three hours per day (dark grey-dashed line). There are basically no differences in the effect between one and two hours of care as a definition. As regards three hours of care we find a considerably stronger short-term effect with a reduction of MCS by 31 per cent of a sd. This probably reflects a higher burden of higher care intensities. Subsequently, however, the effect does not remain on this high level. It immediately drops back to regions similar to those for one and two hours. Most notably, the qualitative result of a considerable short-term effect and a much smaller medium-term effect remains unchanged regardless of the care intensity.[18]



(a) Different care intensities (b) Control group of never carers

Figure 2.4: Alternative definitions of treatment and control groups (MCS only)
Source: SOEP. Own calculations. Note: The dotted lines indicate 95% confidence bands.

The definition of treatment and control group only in $t = 0$ allows for cases where individuals in the control group start to provide care in later years. This is in fact the case for some 15% of all observations in the control group. It might be suspected that these individuals suffer from a short-term mental health drop later which, compared with the effects in the treatment group, lead to the observed relative decline in the mental health drop of the treatment group. In order to test if this drives the results, we exclude all individuals from the control group that provided care in any year between $t = 1$ and $t = 7$. That is, we only use individuals in the lowest path in Figure 2.2b. In principle, this is not a desirable specification as it bases the control group definition on later outcomes. Thus, it should only be regarded as a brief check wether these individuals drive the results observed above. Figure 2.4b shows that this is not the case. The results are largely the same.

The results suggest a significant short-term effect of informal care-provision on mental health while there is a smaller and not significant medium-term effect. Given that

[17] In the Supplementary Material, we also report the results for females caring four hours and more. The results are comparable.

[18] Although not shown here, also the PCS results are robust to these different definitions.

the vast majority of individuals provides care for about one to three years, the main pathway of these effects is probably the following one. Contemporaneously, care provision is a mentally burdensome task. The short-term effects are mostly generated by individuals who just stopped to provide care or who are still providing care in $t = 1$. As to be expected, this effect increases in care intensity. Yet, after the care episode ceased, individuals recover and their mental health status approaches former levels.

The short-term effect is not necessarily entirely due to care provision. It might be a joint effect of care and the observation of the decline of a beloved person. As most of the previous literature, we cannot disentangle the family effect from the active caregiving effect. As results of Bobinac et al. (2010) suggest, the overall effect is a mixture of both but a caregiving effect remains after controlling for the family effect. Yet, this does not affect the interpretation of the medium-run effect of almost no mental health impairment a couple of years after care provision. Given that the effect in $t = 7$ is very small, it can be concluded that there is less evidence for a scarring effect of care provision. Moreover, since only a handful individuals in the sample care throughout the entire observation period, this result can apparently not be explained by an adaptation effect of care providers to their new situation.

In Section 2.4 we mentioned that we cannot stratify the analysis with respect to the care recipient as we do not have information on who is being cared. We can, however, approach such an analysis by splitting the sample into caregivers below and above the age of 60. The former group has a higher likelihood to care for a parent while the latter should be more likely to care for a spouse. Note that stronger restrictions such an age cutoff at 70 or groups such as unmarried women with at least one parent alive are hardly feasible due to strongly reduced numbers of observations. Figure 2.5 shows the effect over time for both subgroups. Initially, they coincide nearly perfectly. Five years after care is observed, they deviate from each other. Whereas younger carers drop back almost to the initial level, for older carers the impact on their mental score is even stronger. The results could be interpreted such that the active caregiving effect does not depend on the care recipient. However, a likely family effect might arguably be stronger in case of careprovision for a spouse than for oldest old parents. However, the effects come closer after seven years and due to large confidence bands one should interpret these results cautiously.

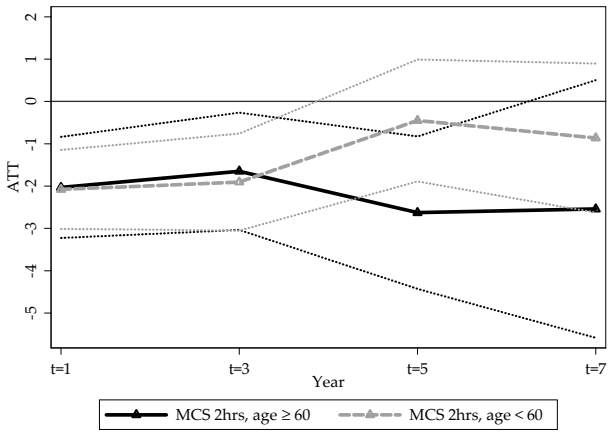


Figure 2.5: Alternative definitions of treatment and control groups (MCS only)
Source: SOEP. Own calculations. Note: The dotted lines indicate 95% confidence bands.

Altogether, the results from this section could be interpreted as good news. While there is a considerable negative short-term effect of contemporaneous caregiving, the scarring effect is less likely to be prevalent. One negative interpretation for these results could, however, be an increased consumption of antidepressants as found by Van Houtven et al. (2005) and Schmitz and Stroka (2013) for the short run. If this would hold for the long run, the mental health score might increase over time due to drug consumption and not due to improved health. Whether this is the case or not requires long-term data on care and drug consumption and is left for future research.

In the Supplementary Material we report results from alternative specifications of the propensity score, the treatment indicator and a subgroup analysis for unmarried women with at least one parent alive who arguably can be identified as caring for their parents.

2.6 Sensitivity analysis

Thus far, we have argued that our estimation strategy allows us to interpret the results in a causal manner since, by fully exploiting the panel information in the SOEP, the conditional independence assumption is likely to hold. However, this inherently untestable assumption might nevertheless fail. For example, in the context of care, it might be particularly challenging to properly control for intrinsic willingness to provide care. Yet, the conditional independence assumption is not necessarily an "all or nothing" assumption and there might be different degrees of its violation. To examine to what extent the magnitude and the significance of our results depend on the potential exclusion of a relevant variable, we follow an approach by Ichino et al. (2008) who refined the suggestions for sensitivity analyses by Rosenbaum and Rubin (1983) and Imbens (2003). This analysis is also in the spirit of the one suggested by Altonji et al. (2005) without the need to make strong parametric assumptions.

Assume that the conditional independence assumption does not hold but that the failure is due to an unobserved variable U. If we could condition on it, we would be able to restore conditional independence:

$$Y_0 \perp\!\!\!\perp T|(X,U).$$

Hence, all the unobserved heterogeneity that results in bias is captured by U. For simplicity, Ichino et al. (2008) follow Rosenbaum and Rubin (1983), who proposed a binary U.

We simulate U by drawing 200 times from the Bernoulli distribution for each individual and estimate the ATT 200 times, conditioning on X as before, but also on U.[19] In simulating U, we make sure that it is both correlated with T and Y such that leaving it out would result in a violation of the conditional independence assumption. Taking

[19]This section contains a non-technical and intuitive discussion of the analysis. A more detailed account is provided in the Supplementary Material published online. For an extensive treatment, refer to Ichino et al. (2008).

the average over all effects provides us with robust point estimates as well as standard errors of the average treatment effect on the treated.[20]

The major question is how strong and in what direction the correlation between U and Y resp. T should be defined. We follow one of the two approaches suggested by Ichino et al. (2008) and set it such that we control the "outcome effect" (own effect of U on Y) and the "selection effect" (effect of U on T). As an illustration, think of U as general intrinsic willingness to provide care: $U = 1$ indicates generally willing, $U = 0$ means not willing. This unobserved variable certainly has a positive selection effect such that willing people are more likely to provide care. It may also have a positive outcome effect if the general willingness is positively correlated with health endowment independent of treatment.

The magnitudes of outcome and selection effects could be arbitrarily chosen. One way to find reasonable values is to use observed binary variables in the data set and calculate the observed selection and the outcome effects of these variables. This gives an indication of the distribution of selection and outcome effects in the data. To bound these effects, one could argue that the unobserved variable U should not have much larger selection and outcome effects than important observed variables, for which we have a long vector, including age, education, and initial health.

We compute these effects for all variables in the sample. Results are reported in the Supplementary Material. We then choose the parameters to simulate U such that is has an effect on treatment and outcome in the same magnitude as the control variable with the highest effect (which is the potential caregiver's age). With these calibrations, no other confounder in the sample (except for age) features such a high effect on mental health and no other makes people select into treatment like the simulated binary confounder U. The first assumed selection effect reflects a positive selection into treatment, i.e., more people with high values of U will take the treatment. Together with a positive outcome effect, we should underestimate effects of care provision on health. The second pair of selection and outcome effects reflects a negative selection into treatment leading to overestimation if this was neglected by the analysis so far.

Figure 2.6 presents the results of both specifications and the baseline specification for MCS. Including a confounder U with characteristics that lead to a positive selection into treatment (the dark grey-dashed line) leads to larger effects of care provision than in the baseline case while we find weaker effects when including a confounder that induces as negative selection (the light grey-dashed line). The lines are parallel-shifted by the confounder. However, in all three cases, we find a significant (both statistically and economically) short-term effect of care provision on mental health which reduces over time. After seven years, the effects are insignificant for most specifications, marginally significant though for the one with a confounder inducing a positive selection into treatment. Thus, if there are further confounders that point in the same direction as most of the variables in our sample, our result will define a lower bound. Furthermore, this would raise the likelihood of significant medium-term effects.

Thus, as long as unobserved effects that are necessarily left out in our analysis do not have a drastically higher impact than observed control variables, we find that the

[20] We use a modified version of the user-written Stata command sensatt (Nannicini, 2007).

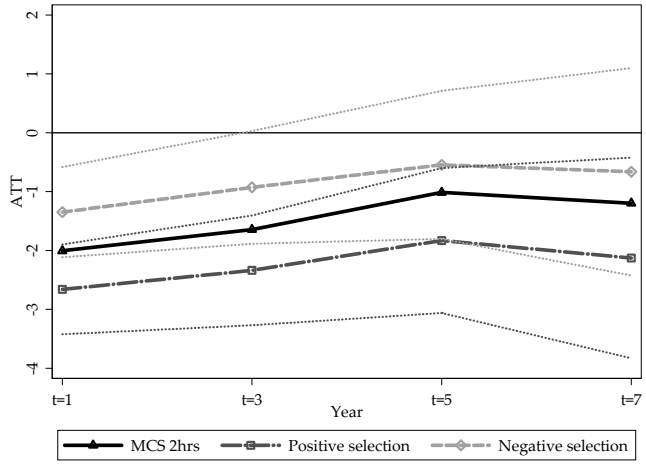


Figure 2.6: Results of the sensitivity analysis (MCS)

Source: SOEP. Own calculations. Note: The dotted lines indicate 95% confidence bands. Strong positive selection assumes a selection effect of $s = 0.25$ and an outcome effect of $d = 0.25$. See the Supplementary Material for exact definitions of s and d and justifications for these values. Strong negative selection assumes a selection effect of $s = -0.25$ and an outcome effect of $d = 0.25$.

average treatment effects we received in the main analysis are robust. Given that we control for a large set of important determinants of care provision, it seems unlikely that there are actually unobserved variables that have such a drastic effect or an effect much stronger than observable covariates.

2.7 Conclusion

This paper examines whether informal care provision affects the report of measures that indicate mental or physical strain among women. We use the German Socio-Economic Panel that identifies informal caregivers by the daily time spent caring. We define caregivers as women who care at least two hours per weekday (but other definitions lead to a similar picture). We evaluate the impact of caregiving on health by help of a regression adjusted matching technique. The problems of unobserved heterogeneity and reverse causality are tackled by exploiting the panel structure of the data set and controlling for pre-treatment outcome as well as stratifying by pre-treatment care status.

While we do not find effects of informal care on physical health in the short- and in the medium-run, our results suggest that there are considerable short-term effects of informal care provision on mental health which, however, attenuate over time. Five years after care provision the effect is still negative but smaller and insignificant. It seems that, contemporaneously, care provision is a mental burden but there is not a large scarring effect. The sensitivity analysis according to Ichino et al. (2008) suggests that these results are stable even for considerable deviations from the conditional independence assumption: the effects are still similar in magnitude even if we falsely have not incorporated a confounder that is stronger than any other one that we have controlled for before.

We contribute to the current debate on how to realign the care system in Germany and countries with similar demographic developments. Our results suggest that there are considerable short-term health effects and although it seems to be good news that the effects are abating, it should not be concluded that there is no need to improve the system as apart from health there are other additional effects of care provision (e.g. labour force participation and wages) that are not analysed here. The current German government put the enhancement of informal care high on the agenda.[21] In particular, the supply of low-threshold services is planned to be expanded and increases in benefits from the long-term care insurance are meant to be spent on those. These services are, e.g., additional help in the household, contact persons in case of any problems, or professional short-term care (also overnight) in case of short-term absence of informal care providers due to sickness, obligations in the job, or holidays. Thus, while family members will certainly continue to play an important role in care provision, these measures are thought to assist them and to reduce the most stressful aspects of care.

The measured effect in this study is an average effect over different groups of care providers. Schmitz and Stroka (2013), for instance, focus on individuals who not only provide informal care but also work full-time. This double burden might well also have health effects in the longer run. This question is left for future research. The main limitations in this study arise from the imperfect data set. Both measures of care provision as well as health indicators are self-reported and potentially measured with error. We do not observe any characteristics of the care recipient. Hence, we cannot distinguish between the family effect that occurs just because a close relative is in need of care and the caregiving effect. However, this should not qualitatively affect the interpretation of the already small medium-term effect. Likewise, as it is not observed whether care recipients receive additional professional care or only informal care, we cannot discriminate between cases in which the caregiver assists professional care and in which she is the only care provider. Moreover, due to data restrictions we are not able to identify the cumulative effect of care provision for many consecutive years. This might go along with even long-run health impairments. However, our representative data suggest that only a very small group of women is faced with the need (and willingness) to provide care for many consecutive years. Moreover, we argue that our approach allows us to answer a question that is more relevant from an individual perspective: if I provide care today, what is my expected health outcome in seven years (irrespective of future events that I cannot control today).

2.8 Appendix

[21] http://www.bmg.bund.de/ministerium/presse/english-version.html

Table 2.5: Table of results

	t=1	t=3	t=5	t=7
Care 2hrs per day (baseline)	-2.00^{***}	-1.64^{***}	-1.01	-1.19
	(0.39)	(0.47)	(0.62)	(0.86)
...care starters	-2.03^{***}	-1.67^{***}	-1.02	-1.21
	(0.40)	(0.49)	(0.64)	(0.88)
...continued care	-1.42^{**}	-0.93^{*}	-0.70	-0.94
	(0.61)	(0.74)	(0.89)	(1.51)
Care 3 hrs per day	-3.02^{***}	-1.44^{**}	-1.00	-1.64
	(0.53)	(0.68)	(0.78)	(1.12)
Care 1 hr per day	-1.90^{***}	-1.59^{***}	-0.48	-0.97
	(0.31)	(0.39)	(0.48)	(0.67)
Observations	28,622	20,288	12,254	5,552
Only never carers in control group:				
Care 2 hrs per day	-2.08^{***}	-1.56^{***}	-1.25^{**}	-0.787
	(0.27)	(0.34)	(0.46)	(0.69)
Observations:	25,914	18,464	11,301	5,166
PCS as outcome:				
Care 2hrs per day	0.14	0.14	0.08	-1.01
	(0.33)	(0.40)	(0.49)	(0.84)

Source: SOEP, own calculations. Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$ indicate the corresponding significance level. Standard errors are in parantheses.

Table 2.6: SF-12v2 questionnaire in the SOEP

	Very Good	Good	Satis-factory	Poor	Bad
How would you describe your current health?					
	Greatly	Slightly	Not at all	–	–
When you ascend stairs, i.e. go up several floors on foot: Does your state of health affect you greatly, slightly or not at all?					
And what about having to cope with other tiring everyday tasks, i.e. where one has to lift something heavy or where one requires agility: Does your state of health affect you greatly, slightly or not at all?					
Please think about the last four weeks. How often did it occur within this period of time, . . .	Always	Often	Some-times	Almost never	Never
◇ that you felt rushed or pressed for time?					
◇ that you felt run-down and melancholy?					
◇ that you felt relaxed and well-balanced?					
◇ that you used up a lot of energy?					
◇ that you had strong physical pains?					
◇ that due to physical health problems					
. . . you achieved less than you wanted to at work or in everyday tasks?					
. . . you were limited in some form at work or in everyday tasks?					
◇ that due to mental health or emotional problems					
. . . you achieved less than you wanted to at work or in everyday tasks?					
. . . you carried out your work or everyday tasks less thoroughly than usual?					
◇ that due to physical or mental health problems you were limited socially, i.e. in contact with friends, acquaintances or relatives?					

Note. Source: SOEP Individual question form. Available at http://panel.gsoep.de/soepinfo2008/.

Supplementary Material

1. Validity of the time allocation measure used for the caregiving definition

As the caregiving definition is based on the self-reported time-use questionnaire of the SOEP, it could be prone to recall bias. For this paper, it is crucial that firstly, individuals do not mix up casual 'care' activities such as taking coffee with their parents with serious care and that secondly, the recalled time allocation has a high degree of valid correlation with the correct time allocation. Sonnenberg et al. (2012) address the latter issue carefully by comparting results from standard survey questions and mobile-phone based experience sampling technology. Their findings indicate a high correlation between both measures, respectively for long-lasting, day structuring activities. Although sporadic activities were also shown to exhibit a high amount of valid correlation, we argue that daily informal care is a day-structuring activity with a long duration. In the loosest definition, we consider individuals as caregivers if their daily time allocation exceeds two hours. This is almost certainly rules out casual care activities. In addition, Van den Berg and Spauwen (2006) also validated time-use questions and find that survey questions on weekly aggregated survey questions tend to underestimate the actual time use (Van den Berg and Pinger, 2014). Hence, there is no major reason to doubt the validity of the time allocation measure in our analysis in general.

2. Other Robustness Checks

Results of alternative specifications

In this section, we change the propensity score specification in two dimensions. First, we do not exclude variables that are insignificant in the first-stage probit regression. Second, we include further variables (the body mass index, a dummy that indicates if the individual smokes, the number of brothers) and, in addition, the functional form of the age variable is changed. In order to assess the robustness of the functional form, the linear and a quadratic term in age are replaced by a set of mutual exclusive dummy variables on the age interval from 20 to 82. Table 2.7 reports how well the control group is balanced to the treatment group.

The result for the alternating specification of the propensity score is depicted in Figure 2.7. It indicates robustness with repect to the functional form specification. The lines for both specifications coincide nearly perfectly. A deviation is first visible five years after treatment where the observations become more scarce. One the one hand this invariance is due to the large set of covariates that we have already conditioned on before. Adding a smoker dummy and the body mass index as a more objective health measure does not provide a lot of new information. On the other hand, for

Table 2.7

	Treated		Controls		Matched controls		Standardized bias		
	mean	sd	mean	sd	mean	sd	unmatched sample	matched sample (0.06)	(0.03)
Stage i): care obligations									
Age of mother									
∈ [30, 39]	.01	.09	.02	.16	.01	.12	-13.35	-5.17	-2.35
∈ [40, 49]	.03	.18	.1	.3	.06	.23	-26.94	-9.21	-3.09
∈ [50, 59]	.08	.27	.13	.34	.1	.29	-18.2	-6.29	-2.19
∈ [60, 69]	.12	.32	.12	.32	.11	.32	.28	.94	.9
∈ [70, 79]	.09	.28	.06	.24	.08	.27	9.18	3.78	1.56
∈ [80, 89]	.09	.28	.02	.14	.07	.25	30.46	7.45	2.56
∈ [90, 99]	.01	.1	0	.03	.01	.09	12.19	2.65	1.27
Mother alive	.46	.5	.48	.5	.47	.5	-5.77	-2.98	-.72
Age of father									
∈ [30, 39]	0	.07	.01	.09	.01	.08	-5.17	-2.24	-1.19
∈ [40, 49]	.02	.12	.08	.27	.04	.19	-30.31	-11.01	-4.47
∈ [50, 59]	.04	.2	.1	.29	.06	.24	-20.71	-6.83	-1.94
∈ [60, 69]	.07	.25	.08	.27	.07	.26	-6.43	-2.13	-.7
∈ [70, 79]	.04	.19	.04	.19	.04	.19	-.35	-.71	-.91
∈ [80, 89]	.01	.11	.01	.1	.01	.11	2.32	.34	-.27
∈ [90, 99]	0	.04	0	0	0	0	5.92	5.92	5.01
Father alive	.19	.39	.34	.47	.24	.43	-32.67	-11.37	4.09
Number of sisters	1.08	1.21	1.09	1.21	1.09	1.22	-1	-1.26	-1.29
Number of brothers	.19	.62	.19	.6	.2	.61	.44	-.38	-.33
.68									
Partner	.81	.39	.68	.47	.77	.42	30.1	9.68	2.54
Age partner	47.77	25.99	35.52	27.08	43.66	26.66	46.15	15.48	4.72
Stage ii): willingness to provide care									
NEURO	4.53	.67	4.37	.72	4.48	.72	22.68	7.25	2.03
CONSC	6.04	.74	5.97	.79	6.02	.77	9.7	2.6	.13
AGREE	5.61	.83	5.58	.84	5.6	.84	3.47	1.38	.79
OPENN	4.37	1.15	4.51	1.12	4.42	1.13	-11.81	-4.5	-2.22
EXTRA	5.02	.91	5.04	.95	5.03	.95	-2.12	-.74	-.27
Positive reciprocity	5.66	.95	5.55	.99	5.64	.97	11.33	2.72	-.53
Negative reciprocity	2.71	1.19	2.87	1.23	2.76	1.22	-12.63	-3.83	-1.05
Acceptance of private funding	3.31	.81	3.29	.8	3.31	.81	2.74	.65	-.04
Age									
∈ [20, 22]	.01	.08	.02	.15	.01	.11	-13.92	-5.36	-2.37
∈ [23, 25]	.01	.07	.03	.18	.02	.12	-20.85	-7.65	-3.26
∈ [26, 28]	.01	.1	.04	.19	.02	.14	-17.05	-6.17	-2.41
∈ [29, 31]	.01	.11	.04	.2	.02	.15	-17.73	-6.33	-2.38
∈ [32, 34]	.03	.16	.05	.22	.03	.18	-12.44	-4.05	-1.08
∈ [35, 37]	.02	.15	.06	.24	.04	.19	-17.93	-6.33	-2.25
∈ [38, 40]	.03	.18	.07	.25	.05	.21	-15.63	-5.48	-1.89
∈ [41, 43]	.04	.2	.07	.25	.05	.22	-10.83	-3.85	-1.43
∈ [44, 46]	.06	.24	.07	.25	.06	.24	-3.63	-.88	.02
∈ [47, 49]	.06	.24	.06	.25	.06	.24	-.63	.12	.28
∈ [50, 52]	.08	.27	.06	.24	.07	.26	8.31	2.99	.87
∈ [53, 55]	.09	.28	.05	.23	.08	.27	12.79	4.29	1.46
∈ [56, 58]	.09	.29	.05	.21	.08	.27	17.34	5.25	1.18
∈ [59, 61]	.09	.29	.05	.21	.07	.26	17.01	5.71	2.34
∈ [62, 64]	.09	.29	.05	.22	.08	.27	15.48	4.91	1.29
∈ [65, 67]	.09	.28	.05	.22	.07	.26	13.09	4.68	1.72
∈ [68, 70]	.06	.24	.04	.21	.05	.23	6.51	2.32	.85
∈ [71, 73]	.05	.22	.03	.18	.04	.2	8.14	2.74	.63
∈ [74, 76]	.04	.2	.03	.16	.03	.18	7.38	3.7	2.13
∈ [77, 79]	.01	.12	.02	.15	.02	.13	-5.82	-1.75	-.36
∈ [80, 82]	.01	.11	.02	.13	.01	.11	-4.6	-1.05	.15
Married	.8	.4	.63	.48	.75	.44	39.1	12.89	3.82
Divorced	.07	.25	.09	.28	.07	.26	-7.57	-2.64	-1.05
Single	.07	.25	.17	.38	.1	.3	-32.78	-10.87	-3.34
Children_hh	.18	.38	.28	.45	.19	.39	-25.65	-8.16	-3.75
Educ general	.14	.34	.17	.38	.14	.35	-9.08	-3.33	-1.76
Educ middle	.57	.5	.49	.5	.56	.5	16.18	5.13	1.7
Foreign	.04	.2	.06	.24	.05	.22	-9.93	-4.05	-2.11
West	.69	.46	.75	.43	.71	.45	-13.53	-4.89	-1.91
Full time	.13	.34	.26	.44	.18	.38	-33.87	-11.75	-4.14
Stage iii): ability to provide care									
MCS	47.38	10.52	49.47	10.12	48	10.82	-20.22	-5.95	-1.4
PCS	46.44	10.02	49.03	10.14	47.36	10.44	-25.64	-9.11	-3.02
Satis health	6.09	2.25	6.53	2.18	6.14	2.26	-19.85	-5.96	-2.03
Satis life	6.19	2.21	6.58	2.17	6.32	2.23	-17.81	-5.88	-.63
BMI	26.25	4.71	25.23	4.75	25.93	4.93	21.58	6.77	1.51
Smoker	.24	.43	.24	.43	.24	.43	-1.55	-.59	-.25
N		1,144		27,406		27,406			

the Kernel-weighting matching method that we have employed, only the rank of the observations with respect to the propensity score is important. This is opposed to inverse probability weigthing where the exact magnitude of the propensity scores affect the results directly.

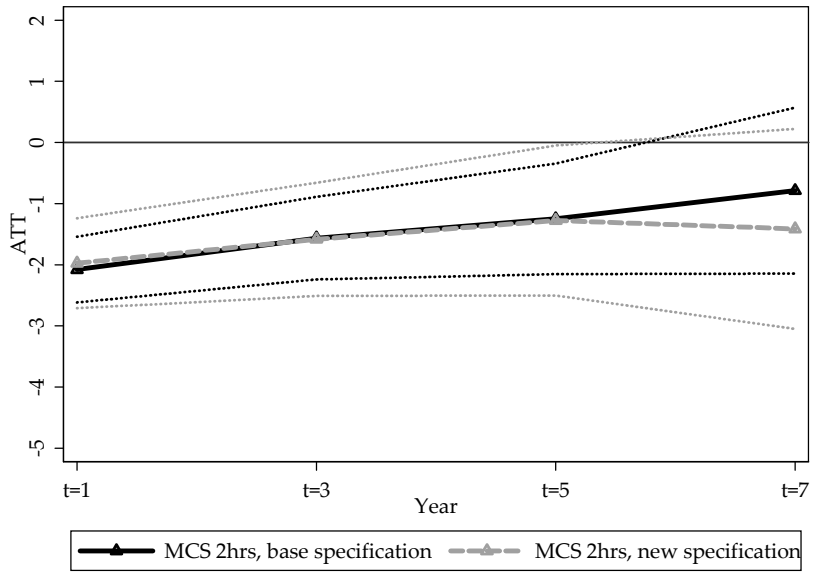


Figure 2.7: Results for an alternative specification of the propensity score vs. basline specification (2hrs MCS only)

Source: SOEP. Own calculations.

Results for different groups of carers

Figure 2.8 illustrates the results for a women that are more likely to care for their parents. To be precise, the sample is conditioned on not being married and having at least one parent alive. The results indicate a similar magnitude of the effects as for the baseline specification. However, due to the imprecise estimates as a result from a strongly reduced sample size, further conclusions cannot be drawn.

Figure 2.9 shows results for an even higher care intensity. In general, the overall picture is not changed. Carers seems to be mentally strained. As time passes, the effect fades out slightly.

Note that for both results, a proper identification with stratification becomes infeasible due to a small sample size.

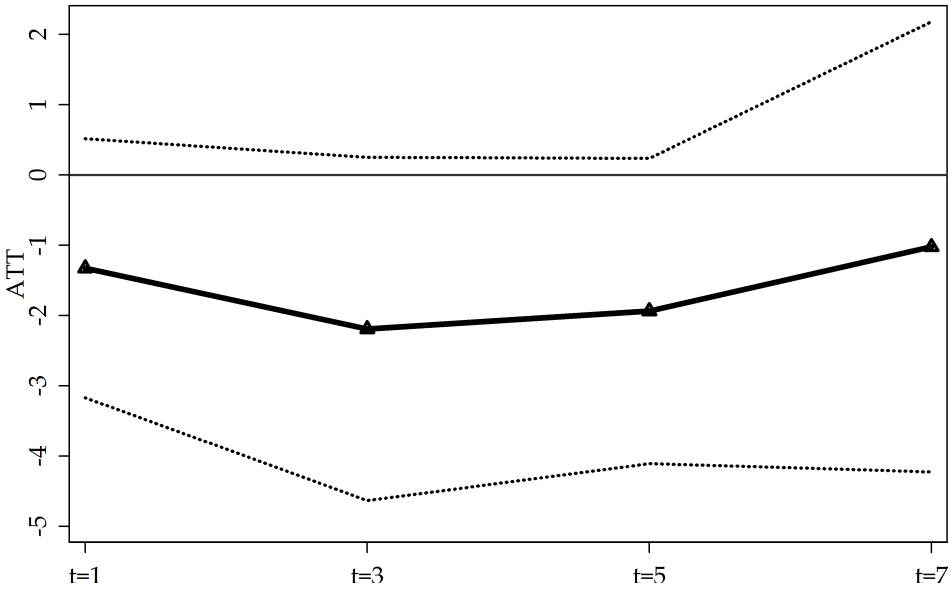
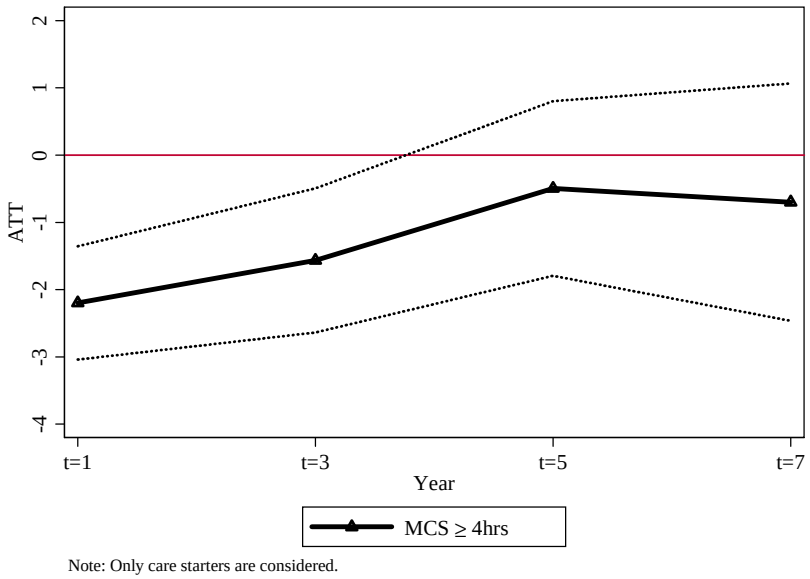


Figure 2.8: Subgroup analysis for unmarried women with at least one parent alive (2 hours, MCS)

Source: SOEP. Own calculations.



Note: Only care starters are considered.

Figure 2.9: Alternative definitions of treatment and control groups (4 hours of care provision, MCS)

Source: SOEP. Own calculations.

3. A Sensitivity Analysis suggested by Ichino et al. (2008)

Assume that the conditional independence assumption does not hold

$$Y_0 \not\perp\!\!\!\perp T|X$$

but that the failure is due to an unobserved variable U. Could we condition on it, we had

$$Y_0 \perp\!\!\!\perp T|(X,U).$$

Hence, all the unobserved heterogeneity that leads to endogeneity problems is captured by U. To keep things as simple as possible, Ichino et al. (2008) follow Rosenbaum and Rubin (1983) who proposed U to be binary. This is appealing, since the distribution of a binary variable is fully determined by its mean. To describe how U affects both treatment and outcome, we define four probabilities p_{ij}, $i \in \{0,1\}; j \in \{0,1\}$ as

$$\begin{aligned} p_{01} &= Pr(U=1|T=0,\hat{Y}=1) \\ p_{00} &= Pr(U=1|T=0,\hat{Y}=0) \\ p_{11} &= Pr(U=1|T=1,\hat{Y}=1) \\ p_{10} &= Pr(U=1|T=1,\hat{Y}=0) \end{aligned} \tag{1}$$

where $\hat{Y} = \begin{cases} 1, & \text{if } Y > \overline{Y} \\ 0, & \text{else} \end{cases}$

and $\overline{Y}$ is the sample mean of Y. Treatment status T and outcome category $\hat{Y}$ are observed in the data and, hence, individuals can be assigned one out of the four probabilities p_{ij} where i denotes treatment status and j indicates whether the outcome exceeds the sample mean. The four above equations fully define the distribution of the hypothetical confounding variable U. Depending on how these probabilities are set by the researcher, the degree of correlation between Y and T varies.

Ichino et al. (2008) define the parameter $s = p_{1.} - p_{0.}$ as the selection effect where

$$p_{i.} = Pr(U=1|T=i) = p_{i0} \cdot P(\hat{Y}=0|T=i) + p_{i1} \cdot P(\hat{Y}=1|T=i) \quad i \in \{0,1\}.$$

The larger this effect, the larger is the effect of U on selection into treatment keeping the outcome fixed. The outcome effect, defined as $d = p_{01} - p_{00}$ reflects the influence of U on the untreated counterfactual outcome. As an example, an outcome effect of $d = 0.1 > 0$ means that the unobserved U positively affects the outcome variables. In the group of non-carers (with p_{0j}), those who are in good health have a higher likelihood of $U = 1$ than those who are in bad health. The higher d the stronger is this correlation. Likewise, a selection effect of $s = 0.1 > 0$ implies that among caregivers the likelihood of $U = 1$ is higher than among non-caregivers. Once we set values for d and s we can derive the four p_{ij} by solving an equation system (as shown below) and simulate U.

Calculation of the p_{ij}.

For an extensive treatment, refer to Ichino et al. (2008). Here we just sketch the idea: Given values for d and s, the four parameters p_{ij} can be derived by solving an equation system of four equations. Assume that $d = 0.1$, $s = 0.1$ and $P(U = 1) = 0.5$. Then we have

$$\begin{aligned} & P(U=1) = 0.5 \qquad (2.1) \\ & = p_{11} * P(\hat{Y}=1|T=1) * P(T=1) + p_{10} * P(\hat{Y}=0|T=1) + P(T=1) \\ & + p_{01} + P(\hat{Y}=1|T=0) * P(T=0) + p_{00} * P(\hat{Y}=0|T=0) * P(T=0) + p_{00} * P(T=0) \end{aligned}$$

$$p_{11} - p_{10} = 0 \qquad (2.2)$$

$$d = p_{01} - p_{00} = 0.1 \qquad (2.3)$$

$$\begin{aligned} & s = p_{1.} - p_{0.} = 0.1 \qquad (2.4) \\ & = p_{10} * P(Y=0|T=1) + p_{11} * P(Y=1|T=1) \\ & - p_{00} * P(Y=0|T=0) + p_{01} * P(Y=1|T=0) \end{aligned}$$

Choice of d and s

In principle, d and s could be arbitrarily chosen. One way to find reasonable values is to go back to the equation system and – starting the other way around – use observed binary variables in the data set, substitute them for the unobserved U and calculate the selection and the outcome effect of these variables. We compute these effects for all variables in the sample. Results are reported in Table 2.9. We see that most of the variables have selection and outcome effects of at most 0.1 in absolute values. In line with Table 2.4, the control variables with the strongest impacts are age and being married exhibiting selection and outcome effects of up to 0.24. Hence, we argue that given the unobserved variable has an effect on treatment and outcome in the same magnitude as important control variables with the highest effects, parameterisations of $s = 0.25$ and $d = 0.25$ or $s = -0.25$ and $d = 0.25$ are reasonable values.

Assume that the conditional independence assumption does not hold

$$Y_0 \not\perp\!\!\!\perp T|X$$

but that the failure is due to an unobserved variable U. Could we condition on it, we had

$$Y_0 \perp\!\!\!\perp T|(X,U).$$

Hence, all the unobserved heterogeneity that leads to endogeneity problems is captured by U. To keep things as simple as possible, Ichino et al. (2008) follow Rosenbaum and Rubin (1983) who proposed U to be binary. This is appealing, since the

distribution of a binary variable is fully determined by its mean. To describe how U affects both treatment and outcome, we define four probabilities p_{ij}, $i \in \{0,1\}; j \in \{0,1\}$ as

$$\begin{aligned} p_{01} &= Pr(U=1|T=0,\hat{Y}=1) \\ p_{00} &= Pr(U=1|T=0,\hat{Y}=0) \\ p_{11} &= Pr(U=1|T=1,\hat{Y}=1) \\ p_{10} &= Pr(U=1|T=1,\hat{Y}=0) \end{aligned} \tag{1}$$

where $\hat{Y} = \begin{cases} 1, & \text{if } Y > \overline{Y} \\ 0, & \text{else} \end{cases}$

and $\overline{Y}$ is the sample mean of Y. Treatment status T and outcome category $\hat{Y}$ are observed in the data and, hence, individuals can be assigned one out of the four probabilities p_{ij} where i denotes treatment status and j indicates whether the outcome exceeds the sample mean. The four above equations fully define the distribution of the hypothetical confounding variable U. Depending on how these probabilities are set by the researcher, the degree of correlation between Y and T varies.

Ichino et al. (2008) define the parameter $s = p_{1.} - p_{0.}$ as the selection effect where

$$p_{i.} = Pr(U=1|T=i) = p_{i0} \cdot P(\hat{Y}=0|T=i) + p_{i1} \cdot P(\hat{Y}=1|T=i) \quad i \in \{0,1\}.$$

The larger this effect, the larger is the effect of U on selection into treatment keeping the outcome fixed. The outcome effect, defined as $d = p_{01} - p_{00}$ reflects the influence of U on the untreated counterfactual outcome. As an example, an outcome effect of $d = 0.1 > 0$ means that the unobserved U positively affects the outcome variables. In the group of non-carers (with p_{0j}), those who are in good health have a higher likelihood of $U = 1$ than those who are in bad health. The higher d the stronger is this correlation. Likewise, a selection effect of $s = 0.1 > 0$ implies that among caregivers the likelihood of $U = 1$ is higher than among non-caregivers. Once we set values for d and s we can derive the four p_{ij} by solving an equation system (as shown below) and simulate U.

Calculation of the p_{ij}.

For an extensive treatment, refer to Ichino et al. (2008). Here we just sketch the idea: Given values for d and s, the four parameters p_{ij} can be derived by solving an equation system of four equations. Assume that $d = 0.1$, $s = 0.1$ and $P(U=1) = 0.5$. Then we have

$$\begin{aligned} P(U=1) &= 0.5 \\ &= p_{11} * P(\hat{Y}=1|T=1) * P(T=1) + p_{10} * P(\hat{Y}=0|T=1) + P(T=1) \\ &+ p_{01} + P(\hat{Y}=1|T=0) * P(T=0) + p_{00} * P(\hat{Y}=0|T=0) * P(T=0) + p_{00} * P(T=0) \end{aligned} \tag{2.1}$$

$$p_{11} - p_{10} = 0 \tag{2.2}$$

$$d = p_{01} - p_{00} = 0.1 \tag{2.3}$$

$$\begin{aligned} s &= p_{1.} - p_{0.} = 0.1 \\ &= p_{10} * P(Y = 0|T = 1) + p_{11} * P(Y = 1|T = 1) \\ &- p_{00} * P(Y = 0|T = 0) + p_{01} * P(Y = 1|T = 0) \end{aligned} \tag{2.4}$$

Choice of d and s

In principle, d and s could be arbitrarily chosen. One way to find reasonable values is to go back to the equation system and – starting the other way around – use observed binary variables in the data set, substitute them for the unobserved U and calculate the selection and the outcome effect of these variables. We compute these effects for all variables in the sample. Results are reported in Table 2.9. We see that most of the variables have selection and outcome effects of at most 0.1 in absolute values. In line with Table 2.4, the control variables with the strongest impacts are age and being married exhibiting selection and outcome effects of up to 0.24. Hence, we argue that given the unobserved variable has an effect on treatment and outcome in the same magnitude as important control variables with the highest effects, parameterisations of $s = 0.25$ and $d = 0.25$ or $s = -0.25$ and $d = 0.25$ are reasonable values.

Table 2.8: Distribution of p_{ij} across control variables in the sample

	p01	p00	d	p1.	p0.	s	Effect
Stage i): care obligations							
Age of mother							
$\in [30, 39]$	0.01	0.03	-0.02	0.01	0.02	-0.01	(+)
$\in [40, 49]$	0.04	0.12	-0.08	0.04	0.1	-0.06	(+)
$\in [50, 59]$	0.08	0.14	-0.06	0.07	0.13	-0.06	(+)
$\in [60, 69]$	0.11	0.12	-0.01	0.12	0.12	0	$(\mp)$
$\in [70, 79]$	0.07	0.05	0.02	0.09	0.06	0.03	(+)
$\in [80, 89]$	0.05	0.02	0.03	0.08	0.02	0.06	(+)
$\in [90, 99]$	0.01	0	0.01	0.01	0	0.01	(+)
Mother alive	0.41	0.5	-0.09	0.46	0.48	-0.03	(+)
Age of father							
$\in [30, 39]$	0	0.01	-0.01	0	0.01	0	$(\mp)$
$\in [40, 49]$	0.02	0.09	-0.07	0.02	0.08	-0.06	(+)
$\in [50, 59]$	0.05	0.11	-0.06	0.04	0.10	-0.06	(+)
$\in [60, 69]$	0.07	0.08	-0.01	0.06	0.08	-0.02	(+)
$\in [70, 79]$	0.03	0.03	0	0.04	0.04	0	$(\mp)$
$\in [80, 89]$	0.01	0.01	0	0.01	0.01	0	$(\mp)$
$\in [90, 99]$	0	0	0	0	0	0	$(\mp)$
Father alive	0.20	0.35	-0.15	0.20	0.34	-0.14	(+)
Number of sisters	0.30	0.26	0.04	0.28	0.26	0.02	(+)
Partner existent	0.81	0.66	0.15	0.80	0.67	0.13	(+)
Age of partner	0.78	0.56	0.22	0.77	0.59	0.18	(+)
Stage ii): willingness to provide care							
NEURO	0.64	0.55	0.09	0.56	0.47	0.09	(+)
CONSC	0.59	0.53	0.06	0.63	0.59	0.04	(+)
AGREE	0.49	0.46	0.03	0.52	0.53	-0.01	(-)
OPENN	0.44	0.49	-0.05	0.46	0.52	-0.06	(+)
EXTRA	0.41	0.40	0.01	0.47	0.50	-0.03	(-)
Positive reciprocity	0.58	0.48	0.10	0.55	0.50	0.05	(+)
Negative reciprocity	0.48	0.50	-0.02	0.43	0.46	-0.03	(+)
Acceptance of private funding	0.53	0.49	0.04	0.50	0.48	0.02	(+)
Age	0.7	0.44	0.26	0.71	0.48	0.24	(+)
Age squared	0.62	0.38	0.24	0.63	0.42	0.21	(+)
Married	0.82	0.60	0.22	0.80	0.63	0.17	(+)
Divorced	0.06	0.09	-0.03	0.06	0.09	-0.03	(+)
Single	0.07	0.19	-0.12	0.07	0.17	-0.10	(+)
Children in hh	0.17	0.32	-0.15	0.18	0.30	-0.12	(+)
Educ gen	0.19	0.19	0	0.17	0.18	-0.01	$(\mp)$
Educ middle	0.53	0.48	0.05	0.54	0.49	0.05	(+)
Foreign	0.05	0.07	-0.02	0.04	0.06	-0.02	(+)
West	0.68	0.72	-0.04	0.68	0.75	-0.07	(+)
Full time	0.15	0.27	-0.12	0.13	0.26	-0.13	(+)
Stage iii): ability to provide care							
MCS	0.30	0.34	-0.04	0.45	0.55	-0.10	(+)
PCS	0.40	0.52	-0.12	0.44	0.57	-0.13	(+)
Satisfaction health	0.40	0.48	-0.08	0.48	0.58	-0.10	(+)
Satisfaction life	0.48	0.55	-0.08	0.57	0.68	-0.10	(+)

All variables are transformed into binary indicators where the threshold is the sample average. Note: (+) means an amplifying effect, whereas (-) means that the effect attenuates. $\mp$ indicates no clear effect.

Table 2.9: Distribution of p_{ij} across control variables in the sample

	p01	p00	d	p1.	p0.	s	Effect
Stage i): care obligations							
Age of mother							
$\in [30,39]$	0.01	0.03	-0.02	0.01	0.02	-0.01	(+)
$\in [40,49]$	0.04	0.12	-0.08	0.04	0.1	-0.06	(+)
$\in [50,59]$	0.08	0.14	-0.06	0.07	0.13	-0.06	(+)
$\in [60,69]$	0.11	0.12	-0.01	0.12	0.12	0	$(\mp)$
$\in [70,79]$	0.07	0.05	0.02	0.09	0.06	0.03	(+)
$\in [80,89]$	0.05	0.02	0.03	0.08	0.02	0.06	(+)
$\in [90,99]$	0.01	0	0.01	0.01	0	0.01	(+)
Mother alive	0.41	0.5	-0.09	0.46	0.48	-0.03	(+)
Age of father							
$\in [30,39]$	0	0.01	-0.01	0	0.01	0	$(\mp)$
$\in [40,49]$	0.02	0.09	-0.07	0.02	0.08	-0.06	(+)
$\in [50,59]$	0.05	0.11	-0.06	0.04	0.10	-0.06	(+)
$\in [60,69]$	0.07	0.08	-0.01	0.06	0.08	-0.02	(+)
$\in [70,79]$	0.03	0.03	0	0.04	0.04	0	$(\mp)$
$\in [80,89]$	0.01	0.01	0	0.01	0.01	0	$(\mp)$
$\in [90,99]$	0	0	0	0	0	0	$(\mp)$
Father alive	0.20	0.35	-0.15	0.20	0.34	-0.14	(+)
Number of sisters	0.30	0.26	0.04	0.28	0.26	0.02	(+)
Partner existent	0.81	0.66	0.15	0.80	0.67	0.13	(+)
Age of partner	0.78	0.56	0.22	0.77	0.59	0.18	(+)
Stage ii): willingness to provide care							
NEURO	0.64	0.55	0.09	0.56	0.47	0.09	(+)
CONSC	0.59	0.53	0.06	0.63	0.59	0.04	(+)
AGREE	0.49	0.46	0.03	0.52	0.53	-0.01	(-)
OPENN	0.44	0.49	-0.05	0.46	0.52	-0.06	(+)
EXTRA	0.41	0.40	0.01	0.47	0.50	-0.03	(-)
Positive reciprocity	0.58	0.48	0.10	0.55	0.50	0.05	(+)
Negative reciprocity	0.48	0.50	-0.02	0.43	0.46	-0.03	(+)
Acceptance of private funding	0.53	0.49	0.04	0.50	0.48	0.02	(+)
Age	0.7	0.44	0.26	0.71	0.48	0.24	(+)
Age squared	0.62	0.38	0.24	0.63	0.42	0.21	(+)
Married	0.82	0.60	0.22	0.80	0.63	0.17	(+)
Divorced	0.06	0.09	-0.03	0.06	0.09	-0.03	(+)
Single	0.07	0.19	-0.12	0.07	0.17	-0.10	(+)
Children in hh	0.17	0.32	-0.15	0.18	0.30	-0.12	(+)
Educ gen	0.19	0.19	0	0.17	0.18	-0.01	$(\mp)$
Educ middle	0.53	0.48	0.05	0.54	0.49	0.05	(+)
Foreign	0.05	0.07	-0.02	0.04	0.06	-0.02	(+)
West	0.68	0.72	-0.04	0.68	0.75	-0.07	(+)
Full time	0.15	0.27	-0.12	0.13	0.26	-0.13	(+)
Stage iii): ability to provide care							
MCS	0.30	0.34	-0.04	0.45	0.55	-0.10	(+)
PCS	0.40	0.52	-0.12	0.44	0.57	-0.13	(+)
Satisfaction health	0.40	0.48	-0.08	0.48	0.58	-0.10	(+)
Satisfaction life	0.48	0.55	-0.08	0.57	0.68	-0.10	(+)

All variables are transformed into binary indicators where the threshold is the sample average. Note: (+) means an amplifying effect, whereas (-) means that the effect attenuates. $\mp$ indicates no clear effect.

Chapter 3

Informal care and long-term labor market outcomes[1]

3.1 Introduction

The effects of informal care provision on caregiver's labor force outcomes have been subject to a large literature in the previous two decades. Labor supply reactions (of females, mostly) have been studied by, e.g., Carmichael and Charles (1998), Heitmueller (2007), Ciani (2012), Casado-Marín et al. (2011), Bolin et al. (2008), Ettner (1995, 1996), Crespo and Mira (2014), Heger (2014), Meng (2012, 2013), where the effects range from small to very large (up to 30 percentage points) reductions in the probability to work for pay.[2] The effect on working hours, as studied by, e.g., Wolf and Soldo (1994), Casado-Marín et al. (2011), Bolin et al. (2008), Ettner (1996), Johnson and Sasso (2000), and Van Houtven et al. (2013) are quite mixed, while wage penalties are more consistently found (Van Houtven et al., 2013, Carmichael and Charles, 2003, Heitmueller and Inglis, 2007).

All of these studies have in common that they look at the contemporaneous effect of caregiving on labor market outcomes. As many societies aim at increasing female labor force participation, one often reported policy implication is to set up more flexible work-arrangements to facilitate informal caregiving while keeping the job (Heitmueller, 2007). However, one might argue that negative short-term effects do not pose severe problems – both for the caregivers and societies as a whole – if they are not persistent. Caregiving spells typically last only a couple of years and as soon as caregivers who put their labor force participation on hiatus return to the labor market after cessation of their caregiving spell, the life-time opportunity costs of caregiving might not be too large. However, caregivers are often in the age of 50+ and might have problems to return into the labor force once they left it – either because they

[1] This paper is written jointly with Hendrik Schmitz and is published as: Schmitz, H., and Westphal, M. (2017). Informal Care and Long-term Labor Market Outcomes. Journal of Health Economics, 56(Supplement C), 1 – 18. Funding by the Fritz-Thyssen Stiftung is gratefulle aknowledged.

[2] Relatedly, Geyer and Korfhage (2015b,a) study incentive effects of the long-term care insurance on care provision and labor tsupply.

voluntarily decide to stay absent or because they cannot return due to labor market frictions.[3] This would imply negative consequences that potentially add up over many years after their caregiving period. Thus, to draw conclusions about the holistic costs of care, it is necessary to turn to a longer-run perspective since the cost of caring might be more complex than forgone income for the time spent caring.

This study looks at longer term labor market effects up to eight years after care provision. Evaluating the persistence of effects is the main contribution of this paper. As far as we are aware, only three papers explicitly move away from the contemporaneous perspective. Fevang et al. (2012) use Norwegian data to study labor market outcomes up to around 10 years before and 5 years after the death of a lone parent and do find notable effects on labor market participation (for women, not for men) around the death which, however, are not persistent. On the other hand, reliance on social assistance increases persistently for men. Although the authors do not observe actual care provision these effects can largely be ascribed to informal care obligations. Skira (2015) explicitly takes into account the dynamic effects on labor supply as one of the first papers in this literature. She estimates a dynamic discrete choice model that is underpinned with a theoretical framework. Her results highlight existing labor market frictions for caregivers as their reduced labor supply in the US due to caregiving persists over time. Michaud et al. (2010) also estimate a structural model in order to look at dynamic effects of caregiving on employment. Yet, the authors do not explicitly look at long-run effects and effects beyond three periods after caregiving are not reported.

We use a representative German data set to assess short- and longer-term effects of care provision on labor market outcomes such as the probability to work full-time, to be in the labor force, the number of weekly working hours (conditional on working) and hourly wages. In Germany, the largest European economy, there were 2.6 million people in need of care in 2013 (Statistisches Bundesamt, 2015) and this number is estimated to increase steadily to outnumber 3.4 million people demanding care services by 2030 (Augurzky et al., 2013). Even according to these official – and probably underestimating – numbers, 1.9 million received care in their private home and 1.3 exclusively received informal care (typically by close relatives) making this the most important pillar of the German long-term care system. Thus, not only due to its size as the largest European labor market, Germany is an interesting country to study: it is rapidly aging and already now has a large informal care sector which is even going to increase in the future.

Apart form the longer-term perspective we, as another contribution to the literature, also take the dynamic nature of caregiving spells into account and use sequential inverse probability weighting (IPW) estimators as suggested by Lechner (2009b) and Lechner and Miquel (2010) to estimate effects of up to three consecutive years of care provision. A further, if minor, contribution comes from the methodological side where we offer an identification strategy that relies on less functional form assumptions than the previous literature on short-run effects but also than Skira (2015) and Michaud et al. (2010).[4] Our strategy rests on (sequential) conditional independence

[3] The same holds for switching from full-time to part-time work.

[4] Certainly, this is not to say that we make less or weaker assumptions than the previous literature in general, merely that we make different ones, thereby complementing the picture.

assumptions (CIA) which we justify by exploiting cross-sectional but also longitudinal information from our rich household survey, the German Socio-Economic Panel (SOEP). In auxiliary analyses, we relax the CIA to identify effect bounds under weaker assumptions. Other sensitivity tests such as placebo regressions imply that remaining time-invariant unobservables are unlikely to lead to an upward bias of our estimates.

Our main finding is that female caregivers reduce the probability to work full-time by 4 percentage points (at a baseline probability of 35 per cent). The effect is persistent over a period of eight years and seems to be mainly driven by switches to part-time work. High care intensities and longer episodes, however, also increase the long-run probability to leave the labor force. When we move away from point identification to effect bounds, the reduction in full-time work changes to an interval of 2.4 to 5.0 ppts. As another finding, wages seem to be unaffected contemporaneously but are significantly lower 8 years after the start of a care episode.

The paper proceeds as follows. Section 3.2 gives a brief introduction into the German long-term care system. Section 3.3 presents the data and how we exploit the panel structure. Section 3.4 lays out the estimation strategy and reports results of the baseline (static) model. Section 3.5 scrutinizes the identifying assumptions and allows for deviations. Results of the dynamic model are reported in Section 3.6, while some alternative specifications are carried out in Section 3.7. Section 3.8 concludes.

3.2 Institutional background

The German social long-term care insurance system was introduced in 1995 as a pay-as-you-go system.[5] It is financed by a mandatory pay payroll tax deduction of currently 2.35 per cent of gross labour income (2.6 per cent for employees without children). In order to qualify for benefits, individuals need to be officially defined as care recipients and be classified into one of now four care levels. In care level one individuals need support in physical activities for at least 90 minutes per day and household help for several times a week. Individuals in need of more care are classified into care levels two or three, where the benefits increase in care levels. In addition, to acknowledge the care needs of people with dementia, care level 0 has been added in 2013, if they suffer from limited activities of daily living (but do not qualify for one of the other care levels).

Benefits also depend on the type of care, where monthly payments for informal care range from 123€ (level zero) to 244€ (level one) and to 728€ (level three), for professional ambulatory care from 468€ to 1,612€ and for professional nursing home care from 1,064€ to 1,550€ . The latter, in particular, does not fully cover the expenses for nursing home visits and copayments of up to 50 per cent are standard. Copayments for professional ambulatory care are smaller and amount to an average of 247€ or about 20 per cent (Schmidt and Schneekloth, 2011). Social welfare may step in if individuals are not able to bear the copayment. Nevertheless, the decision for formal or informal ambulatory care might also be driven by financial aspects.

[5] This section is taken almost unchanged from Schmitz and Westphal (2015).

The introduction of the insurance system in 1995 stressed the family as the main provider of care, as it is thought to provide care cheaper, more agreeable, and more efficiently. From the care recipient's perspective, the decision to receive informal care typically expresses a preference for being cared by familiar relatives or friends. In some cases, informal care recipients are additionally supported by professional carers. These are, on average older recipients with a higher care level and, thus, a higher care burden (Schulz, 2010). Apart from the care burden, a reason for professional care can be the absence of appropriate informal caregivers, either because they chose to only participate in the labour market or because their own physical or mental health conditions prohibits the full amount of necessary care provision.

From the caregiver's perspective, affection and sense of responsibility towards a loved parent or spouse mainly drive the decision to provide care. Although the insurance benefits for informal care are often passed on to the care provider this comparably small amount cannot be regarded a financial incentive to provide care, as it is also needed to cover other expenses for care provision (see Schmidt and Schneekloth, 2011 for all points). Even if the the caregiver took the benefit fully as a remuneration, the hourly rate would amount to app. the 10% quantile of the female wage distribution. However, the insurance funds do pay pension contributions for informal carers who provide care at least 14 hours a week (Schulz, 2010). In 2002, people cared on average 14 hours per week for care recipients whose assessment of needs is at least classified as the lowest official category (Schneekloth and Leven, 2003).

Between 2001 and 2013 there were only minor adjustments to the German long-term care system. They were minor because benefits were increased but only to keep pace with the inflation (Rothgang, 2010) and, thus, did not change the incentives to provide care. As of 2008, employed individuals are allowed to take a 10 day (not repeatable) unpaid leave to organize or provide care in case of an incidence of care dependency in the family. However, only very few caregivers make use of this.[6] Thus, the tasks of informal caregivers as well as financial incentives remained similar over time.

3.3 Data

3.3.1 Sample selection

We use data from the German Socio-Economic Panel (SOEP) which is an annually repeated representative panel survey on households and persons living in Germany (Wagner et al., 2007). Since 1984 it covers many questions on different life domains such as work, health, time use and education. On average, the survey contains about 22,000 individuals. We use data from the waves 2001 – 2013 as these include information on informal care provision.

Informal care is defined by the answer to the following question, "What is a typical day like for you? How many hours do you spend on care and support for persons in

[6] Schmidt and Schneekloth (2011) report that only 9,000 out of possibly 150,000 made use of this until 2011. The most frequent reason for not making use in their survey was that individuals were not aware of the possibility.

need for care on a typical weekday?".[7] Around 40% of those in the sample who state a positive number report to care for one hour per day. 25% care for two hours and the remaining 35% for three or more hours. Given that this is self-reported information from the time use questionnaire, we collapse this information into a binary variable which should considerably reduce measurement error – individuals are probably much more likely to recall any care provision than the exact number of hours. Our treatment variable D is defined as the indicator for providing care at least one hour per day. Specifications with two or three hours as relevant thresholds are also presented below.

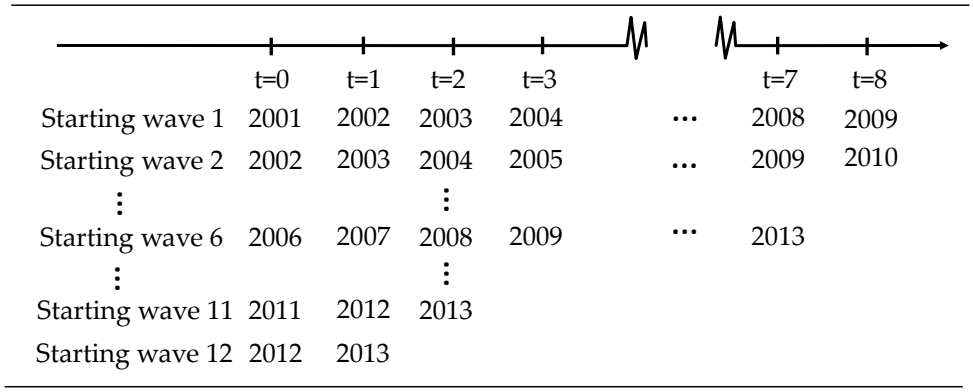


Figure 3.1: Time structure of the data

Source: Own illustration. This figure shows the time structure of the data. Individuals who are observed in the wave of 2001 are called to be from "Starting wave 1". These individuals can, in principle, be observed in all following SOEP waves until the year 2013. Necessary information from future waves are merged to the information from the starting wave. Information of year 2002 is defined to be of year $t = 1$, information of year 2003 is defined to be of year $t = 2$ and so on. Individuals who are observed in the wave of 2002 are called to be from "Starting wave 2". Again, future information is merged, and so on. The same individuals, but at different points in time, can appear in different starting waves.

Figure 3.1 displays the time structure of our data set and shows that we pool observations from different waves. Multiple observations from the same individual are taken into account by using clustered standard errors on individual level in the estimation models. Individuals from the first wave we use (the wave of 2001) can, in principle, be followed for 13 years until 2013. Individuals from wave 11, for instance, can be followed for three years until 2013. We standardize all calendar years across waves to years $t = 0$ to $t = 8$, where $t = 0$ is merely used to define a relevant sample of individuals who did not provide care in this starting period. We drop the years larger than $t = 8$ (relevant only for the first four starting waves) in order to also have a sufficient number of observations to estimate the long-run effects. We restrict the sample to women between 25 and 64 years, since they are in prime working age. Although there is a considerable amount of male informal caregivers in Germany as well – yet, less than female –, we do not carry out an analysis for men. Our estimation procedure below is fairly demanding in terms of sample size (in particular for the long-run effects and the partially dynamic model) and we consider the number of male caregivers in the sample to small for the econometric analysis in this study. We opted

[7] Note that there is another question on child care in the time use questionnaire. Thus, this question explicitly addresses elder care.

against a pooled specification for two reasons. First, we would expect different effects for men and women. Second, more importantly, we assume that the (observed and unobserved) determinants to provide care differ across gender. Thus, we would like to make a conditional independence assumption below only within the gender-subgroups but not across. Finally, we only use individuals with full information in all conditioning variables.

3.3.2 Informal care paths

Table 3.1 reports numbers of observations used in this study. 63,372 person-year observations (from 9,355 different women) meet the sample restrictions (most importantly no care provision in $t = 0$ and age between 25 and 64) of which 2,171 start a care episode in $t = 1$ while 61,201 do not provide care.[8] Among these observations, 16,701 are still in the sample after 8 years (577 of them provided informal care in $t = 1$). This strong reduction has two main reasons: first, recall from Figure 3.1 that individuals who enter the estimation sample from late waves cannot be followed over many years and, second, individuals drop out of the sample if they reach the age of 65. Thus, the long-run effects will be estimated less precisely than the short-run effects.[9]

Table 3.1: Numbers of observation

	Numbers of observations in year							
Care path	1	2	3	4	5	6	7	8
Static model:								
Care in $t = 1$	2,186	2,010	1,760	1,493	1,289	1,044	863	658
No care in $t = 1$	63,121	57,741	50,623	43,649	37,014	30,686	24,957	19,797
Dynamic model:								
Care in $t = 1$ and $t = 2$		838	731	614	529	428	339	260
Care in 1, No care in 2		1,167	1,004	856	743	604	514	391
No care in 1, Care in 2		1,595	1,385	1,191	1,009	843	682	557
No care in 1 and 2		56,449	48,815	42,169	35,772	29,668	24,139	19,138
Care in $t = 1, 2$ and 3			469	380	327	261	208	152
No care in $t = 1, 2$ and 3			47,330	40,365	34,312	28,465	23,149	18,346

Source: SOEP, own calculations. Note that, for example, the sum of individuals in all four paths in year 2 does not equal the sum of individuals from the static perspective in year 2. This is because these figures are based on the estimation samples and due to missing control variables in year 1, an issue that is irrelevant for the static case (explained in Section 3.4) but relevant for the dynamic one (explained in Section 3.6).

Turning to a dynamic perspective, 771 of all women with information in year 2 provide care in both years 1 and 2. More individuals, 1,045, only provided care in $t = 1$ but not in $t = 2$. This reflects the result that most care episodes only last for one year (see below). 222 women who care both in year $t = 1$ and $t = 2$ can be followed until year

[8] These numbers hold for the outcome variables full-time work and labor force participation and are lower for wages (where we have 37,668 person-year observations and hours worked (38,357) as these outcomes are conditional on being employed.

[9] Robustness checks with an age cut-off of 57 that close the second channel for attrition yield the same results, see below.

8. The two bottom lines of Table 3.1 also report numbers of observations for those who cared or did not care in three consecutive years. 417 women are observed to care at least three consecutive years.

Figure 3.2 illustrates the distribution of care durations in our sample. It shows that 60% of all care spells in the sample that start in $t = 1$ last for one period, 18% for two periods, and 7% for three periods. The median care provision duration is one year, while the average is 1.8. Note, however, that these numbers only include spells of consecutive care provision. Interrupted spells like, e.g., care in $t = 1$, no care in $t = 2$, care in $t = 3$, count as duration of one period in Figure 3.2. A potential reason for this interruption could be measurement error in self-reported care provision status, or, less likely, interruptions due to longer hospital or nursing home stays of the care recipient. Most likely, of course, it could also reflect the end of a care episode for a certain care recipient and a later start of a new one for another recipient. As we do not have complete information on the care recipient, this cannot be verified. Figure 3.20 in the supplementary materials shows the same figure for a case where we impute interrupted spells to consecutive spells.[10] This comprehensive change towards longer care spells still results in 75% of all spells lasting for up to three years (mean duration 2.3 years). The averages are in line with those reported in Müller et al. (2010) who use administrative data to estimate that care recipients receive informal care for 2.1 years, on average, in Germany and, thus, show that our care indicator seems to be a useful measure.

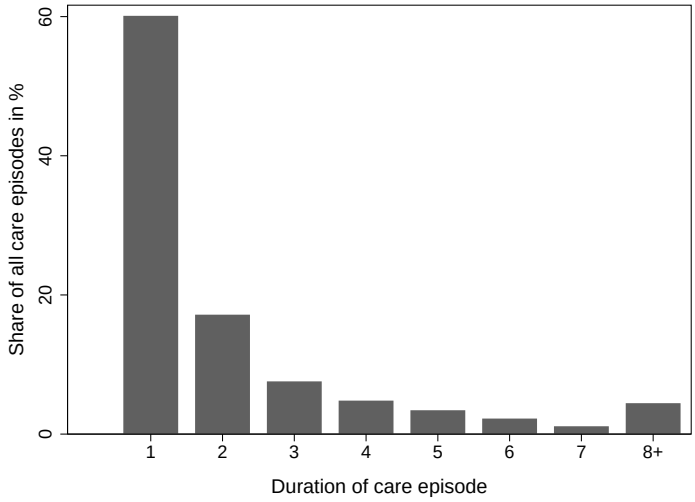


Figure 3.2: The distribution of care spells

Source: SOEP, own calculations. This graph shows the distribution of episodes of consecutive care of at least one hour per day. The data are restricted to starting waves 1 to 5 as defined in Figure 3.1 to ensure that every spell can last for at least 8 years. Note that the vast majority of individuals does not provide care at all (about 90% of all individuals).

[10] In a robustness check we also perform the main analysis using this sample with imputed care spells and the results hardly change.

3.3.3 Outcome variables

We use four different outcome variables: an indicator of full-time work, an indicator of being employed, weekly hours worked (conditional on positive hours) and gross hourly wages (conditional on positive hours). Full-time work and employment are taken from the subfile "generated variables" in the SOEP and are based on a question on current employment status. Being employed means either working full-time, part-time, vocational training, or marginal and irregular part-time employment. Non-employment includes non-working individuals, those in military/community service, maternity leave, and employed persons in a phased retirement scheme whose current actual working hours are zero (SOEP Group, 2014). Working hours are current actual average working hours (including overtime) as reported by the individuals and not contracted working hours. Implausible answers are replaced to missing values by the SOEP group (SOEP Group, 2014). Gross hourly wages are defined as (deflated) gross monthly labor income divided by the product of 4.3 and weekly hours worked.

Table 3.2: Sample means of outcome variables by care status

	Caregivers in $t = 1$	Non-carers in $t = 1$
Full-time, $t = 1$	0.27 (0.44)	0.35 (0.48)
Full-time, $t = 0$	0.31 (0.46)	0.36 (0.48)
Employed, $t = 1$	0.73 (0.45)	0.76 (0.43)
Employed, $t = 0$	0.74 (0.44)	0.77 (0.42)
Hours, $t = 1$ if > 0	31.18 (13.95)	32.31 (13.19)
Hours, $t = 0$ if > 0	32.11 (13.80)	32.43 (13.23)
Hourly wage, $t = 1$	13.98 (8.41)	14.16 (8.79)
Hourly wage, $t = 0$	13.84 (8.00)	14.11 (8.78)

Source: SOEP, own calculations.

Table 3.2 shows sample means of the outcome variables in years 0 and 1 stratified by caregiver status in year 1. Non-carers have higher labor force participation and wages than caregivers. For instance, while the likelihood to work full-time is 35% for non-carers, it is 27% for carers. Somewhat less pronounced, yet significant differences can be found for the other variables. It remains to be seen whether these short-run differences are due to care provision or just reflect different compositions in both groups – and whether they are persistent if they are, at least in part, due to care provision. Obviously the groups of caregivers and non-carers do differ significantly with respect to their labor market attachment even without care provision. The table also shows average pre-treatment outcomes of year 0 – when both groups do not provide care – and only slightly less pronounced differences between both groups can be observed. Thus, it seems central to control for previous outcomes.

3.4 Baseline analysis

3.4.1 Empirical strategy I – A static design

We are interested in the effect of caregiving on labor market outcomes, both contemporaneously and up to eight years later. Figure 3.3 describes the basic design. In period 1, individuals receive the binary treatment D_1 (for all random variables to come, subscripts denote time in years), which could either be care provision ($D_1 = 1$ and a green circle in Figure 3.3) or no care provision ($D_1 = 0$ and a red circle).[11] We restrict the analysis to the subsample of individuals with $D_0 = 0$, that is, those who did not provide care in $t = 0$. We then observe outcomes Y_1 to Y_8. Y_t stands for the four different outcome variables.

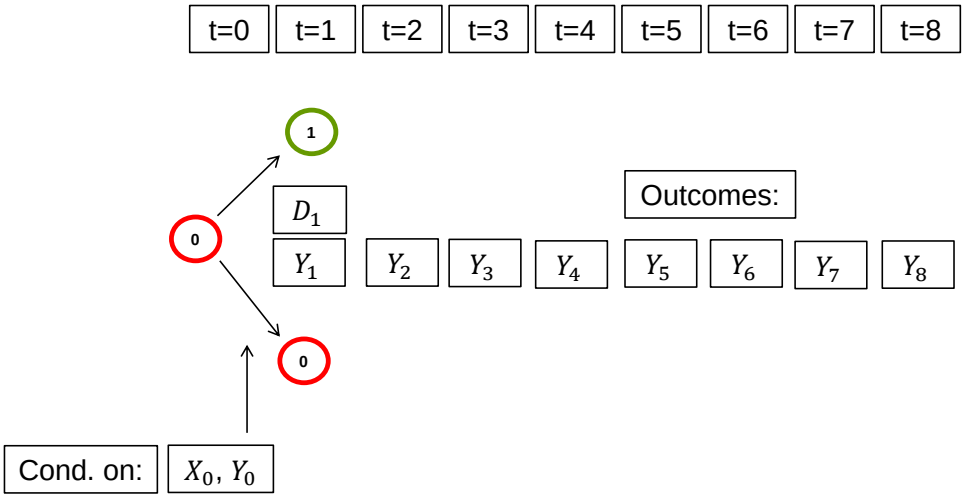


Figure 3.3: Static design

Own illustration.

Each individual has two potential outcomes per period, Y_t^1 and Y_t^0, where the superscripts denote potential outcomes with or without care provision in period 1. The causal effect of providing care in period 1 on labor market outcomes in period t is $Y_t^1 - Y_t^0$. This individual treatment effect is a well-defined parameter but impossible to determine for the researcher as only the factual but not the counterfactual outcome is observed – the observational rule is $Y_t = D_1 Y_t^1 + (1 - D_1) Y_t^0$ if we only define it in terms of D_1 irrespective of any dynamics in D_t over time.

Ideally, we would like to randomly assign individuals to informal care in $t = 1$, follow them over the years and evaluate how they perform on the labor market compared to those who are not assigned to caregiving. This experiment would allow us to assess the average causal response of starting care in $t = 1$ irrespective of how long the person actually provides care. However, we have observational data and care is most naturally a voluntary decision. People individually (potentially altruistically) weight costs and benefits to make a choice that is roughly based on the opportunity,

[11] Note that, for simplicity, we drop subscripts i denoting individuals, such as D_{i1}, throughout the paper.

the willingness, and the ability to provide informal care. While it is not our aim to fully model the decision to provide care, we hope to control for all variables that affect both, treatment and outcomes, leaving the decision to care a random event (conditional on controls). In other words, in order to identify causal effects, we make a conditional independence assumption:

$$Y_t^1, Y_t^0 \perp\!\!\!\perp D_1 | X_0, Y_0 \qquad \forall t > 0$$

Below we go into detail about which variables we account for in X_0 and justify this assumption. We exploit detailed information on individuals' socio-economic background, potential caregiving obligations, health and, most importantly, pre-treatment outcomes Y_0 which capture time-invariant (or permanent) unobserved heterogeneity such as general attitudes towards labor market participation and other hard-to-measure factors such as intrinsic motivation, time preferences, or personality traits. The identifying assumption here is that, conditional on all covariates and past outcomes, the observed treatment is random.[12] We will relax this assumption later in Section 3.5.1.

The parameter we want to estimate is the sample average treatment effect on the treated (ATT), that is, the causal effect of caring for all caregivers in the sample. The identifying assumption enables us to use $E(Y_t^0 | D_1 = 0, X_0, Y_0)$ as a surrogate for the counterfactual $E(Y_t^0 | D_1 = 1, X_0, Y_0)$, and, hence, we overcome the identification problem and can calculate $ATT_t = E(Y_t^1 - Y_t^0 | D_1 = 1, X_0, Y_0)$. We use propensity score kernel matching and inverse probability weighting to achieve this.[13]

Finally, note that the treatment status is only defined in $t = 1$ and not affected by later care provision as this might be endogenously affected by future realizations of the outcome or control variables. This is partly relaxed in a dynamic specification in Section 3.6. In a robustness check of the static version we restrict the control group to individuals that never provide care throughout the full observation period. This cannot be a preferred specification as the definition of the control group depends on future caregiver status. Yet, as the results do not differ compared to the baseline specification (see Figure 3.13 in the Appendix), it makes a strong case that it is no big issue that the control group according to the baseline definition above also includes women who will provide care after $t = 1$.

[12] The set of assumptions is completed by a common support assumption, the stable unit treatment value assumption and the assumption that no control variable is a direct product of the treatment. The first one is naturally fulfilled by restricting the sample to those individuals in the treatment and control group that share a common support of the propensity score. While there is a considerable number of informal care givers, it is fairly low relative to the overall labor force. Thus, regarding the second assumption, we assume the absence of general equilibrium effects and forms of interference between counterfactual outcomes other than the direct treatment effect. The third one is most likely fulfilled by including variables only that are measured one year before the treatment.

[13] As the results hardly differ at all between IPW and Matching, we will only report Matching results for the static version below. A comparison between IPW and Matching is shown in the Appendix in Figure 3.12.

3.4.2 Control variables

The selection of the control variables to estimate the propensity score is crucial in order to make the conditional independence assumptions credible. Here we can make use of the major strength of high quality survey data: the abundance of individual level variables that potentially affect both treatment (paths) and potential outcomes as well as their changes over time. While a drawback compared to many administrative data sets is the comparably small sample size, the advantage is the widespread information on topics such as socio-economic background, health, or preferences that are usually not available in administrative data sets. We consider this crucial for the identifying assumptions.

In deciding for informal care provision one might have three basic blocks of prerequisites in mind. Individuals decide to provide care if (i) they need to, if (ii) they are willing to, and (iii), they are able to provide care. As of (i), individuals are only in the position to decide for care provision if someone close becomes care dependent. We model the intra-social environment by using indicators whether parents are alive, parents' age as well as the number of the potential caregiver's siblings. The latter can reduce the need to provide care for frail parents as siblings could step in.

As of (ii), we select socio-economic characteristics as covariates that also control for the willingness to provide care. This set contains age bin dummies, binary variables on marital status (married, divorced, widowed), whether children live in the household, as well as whether the individual is foreign born. Furthermore, we use character traits measured in the Big Five Inventory (Big5), well-known in psychology for being a proxy of human personality (see McRae and John, 1992 or Dehne and Schupp, 2007) as well as positive and negative reciprocity. The items of the Big5 are: neuroticism, extraversion, openness, agreeableness, conscientiousness.[14] The Big5 are included in the surveys in 2005 and 2009, whereas questions on negative and positive reciprocity are asked in 2005 and 2010. We impute values for the other years.[15]

As of (iii), the own health status determines the ability to provide care. Here, we control for self-rated health, the number of doctor visits in the previous three months and the number of hospital visits in the previous year. Finally, we include pre-treatment outcome variables (that is, previous labor market status), regional dummies for the 16 federal states as well as a full set of year dummies. We would have liked to include more variables which we, however, do not have access to. These are, e.g., the local supply of formal care or characteristics of the care recipient. In order to get as much

[14] More specifically: neuroticism, the tendency of experience negative emotions; extraversion, the tendency to be sociable; openness, the tendency of being imaginable and creative; agreeableness, the dimension of interpersonal relations and conscientiousness, the dimension of being moral and organized (see Budria and Ferrer-i Carbonell, 2012). There are three questions for each of these items which are gathered on a 7-item scale. Although the SOEP captures each item of the Big5 with relatively few questions, surveys revealed sufficient validity and reliability (see Dehne and Schupp, 2007). Furthermore, there is positive reciprocity, the tendency of being cooperative and negative reciprocity, the tendency of being retaliatory.

[15] Specifically, we assign individuals for the years 2001 – 2007 their values of 2005 and for 2008 and later their values of 2009/2010. This assumes stability of personality traits over a short time span which has been empirically confirmed by Cobb-Clark and Schurer (2013), for a different kind of personality trait, however, namely locus of control. Individuals who were neither interviewed in 2005 nor in 2009/2010 are dropped.

as possible out of the observed characteristics and we interact all control variables with each other and include squared terms (as long they are not dummy variables). By capturing potential non-linearities in the determinants to provide care and work, we put less restrictions on the functional form of the outcomes and the propensity score. This, however, results in an extremely large number of control variables that is unfeasible to manage. Thus, we strongly reduce the dimension of the vectors of controls by using Lasso (least absolute shrinkage and selection operator).[16] Specifically, we follow the "double selection"-procedure suggested by Belloni et al. (2014):

1. Select variables (using Lasso) from the full set of (X_0, Y_0) including interactions and squared terms that are relevant to predict D_1.
2. Select variables (using Lasso) from the full set of (X_0, Y_0) including interactions and squared terms that are relevant to predict Y_t.
3. Use the union of variables from steps 1 and 2 as controls.

This procedure works if the "approximate sparsity assumption" (Belloni et al., 2014) holds which, stated verbally, implies the following: the chosen subset using the procedure described above leads to an approximation of the true relationship between outcome and controls, where the approximation error is sufficiently small. Thus, if the CIA holds given the full set of controls and their interactions, it approximately also holds for the chosen subset of controls under the approximate sparsity assumption. All control variables and their sample means are reported in Table 3.4 in the Appendix. An example for the finally chosen ones by the double selection procedure is given in the supplementary materials. This procedure allows to explain a great deal in the variation of the outcome variables by the observed characteristics which could be seen as an indicator for the room for potential failure of the CIA. For instance, a simple regression of full-time work on the chosen controls shows an R^2 of 67%.

Using the union of variables from steps 1 and step 2 as controls might reduce bias but increase the variance if many variables are included that do not affect the outcome (Brookhart et al., 2006). Moreover, we allow variables to be included that have previously been used as instrumental variables (such as the number of sisters or parental age). Using instruments in the propensity score has recently been criticized (Bhattacharya and Vogt, 2012; Wooldridge, 2016). While it is not clear whether these variables are really valid instruments (for instance, Fevang et al., 2012, call previously used instruments in studies on labor market effects of careprovision "potentially invalid"), we also report the main results where only variables from step 2 and no potential instruments are used in the Appendix (Figure 3.15). The results are not very sensitive to this choice. They are exactly the same for conditional hours worked and hourly ages and slightly larger for full-time work and labor force participation. In the main text we opt for reporting the double selection procedure with the more conservative results.

[16] We use the Stata ado lassoShooting provided by Christian Hansen on his website. Of course, any errors are our own responsibility.

3.4.3 Potential failure of the CIA

We can build on a large number of controls to condition on. These include observed preferences as well as predictors of general labor market prospects like age and education. Pre-treatment outcomes should capture time-invariant unobserved factors that might affect both care provision and labor market outcomes. These might, again, be general preferences but also baseline health, such as general unobserved frailty.

Yet, any unobserved changes in factors between our baseline year t=0 and t=1 that affect both, labor supply and informal care would lead to a violation of our identifying assumption because we then would falsely attribute the partial correlation of this unobserved factor between care and labor supply induced by this factor to the effect of informal care. What are such factors? One possibility is, for instance, a shock of mental or physical health that changes one's personal priorities from work to care or vice versa. It is also conceivable that a health shock makes women withdraw from both work and care obligations simultaneously. In the same vein, an unforeseen shock in the opportunity cost to provide care or expectations about the job stability (that arises for instance if the firm where one is employed announces to size down) may also lead to a bias in the causal effect of informal care on labor supply because labor market prospects affect informal care and not vice versa.

A similar reasoning holds for the death of a parent or spouse. This should, on average, reduce the likelihood to provide care, if the parent was a care recipient. Moreover, if there is an own effect of a parent's death on labor force participation, it should be negative. Another problem could be an inadequate measurement of opportunity cost to provide care, general labor market attachment, and expectations about the job stability. Women who recently experienced difficulties in their current job (say, a demotion or missed promotion, for example) or expect to loose their job in the near future might be more willing to provide care.

In the Supplementary Materials we present several analyses to defend our identification strategy. First, as a simple measure, we account for the labor market history (participation and wages) in the five years prior to the start of the care spell and for expectations about job stability. This is not the preferred specification due to loss of numbers of observations. Second, we drop individuals who experienced a health shock or who lost a parent between wave 0 and 1 and delete two potentially important time-varying confounders. Thereby, we see that the health effect on care provision is not very strong and certainly not able to drive the results. Finally, we scrutinize the results with respect to potential systematic measurement error in the treatment variable. More sophisticated, in Section 3.5, we openly allow for additional confounders by simulating them and taking them into account. By this we determine bounds that most likely include the true effects.

3.4.4 Estimation results – Static model

Figure 3.4 reports the effects of caring in year 1 on all outcome variables across time until year 8. These are matching results. Figure 3.11 in the Appendix reports matching quality for full-time work as an outcome variable and shows that covariate balance

is achieved by matching. This is also the case for the other outcomes (not reported). Table 3.5 reports the exact numbers of the estimation results, standard errors, and bandwidths used. Figure 3.12 in the Appendix compares matching results to IPW. The differences are negligible. Therefore, we stick to matching as the method that is more standard in applied econometrics. The upper left panel shows the findings for full-time work. The probability to work full-time is reduced by around 4 percentage points (ppts.) when women start to provide care. This is a considerable effect given an average probability to work full-time of around 35 per cent. Moreover, it persists over the entire observation period, although the confidence bands widen over time due to fewer observations. As only a small fraction of caregivers in $t = 1$ still (or also) provides care in year 6, 7, or 8, this can be interpreted as evidence that some women leave full-time employment due to care obligations and, later, when the care spell ceased, do not return to full-time employment.

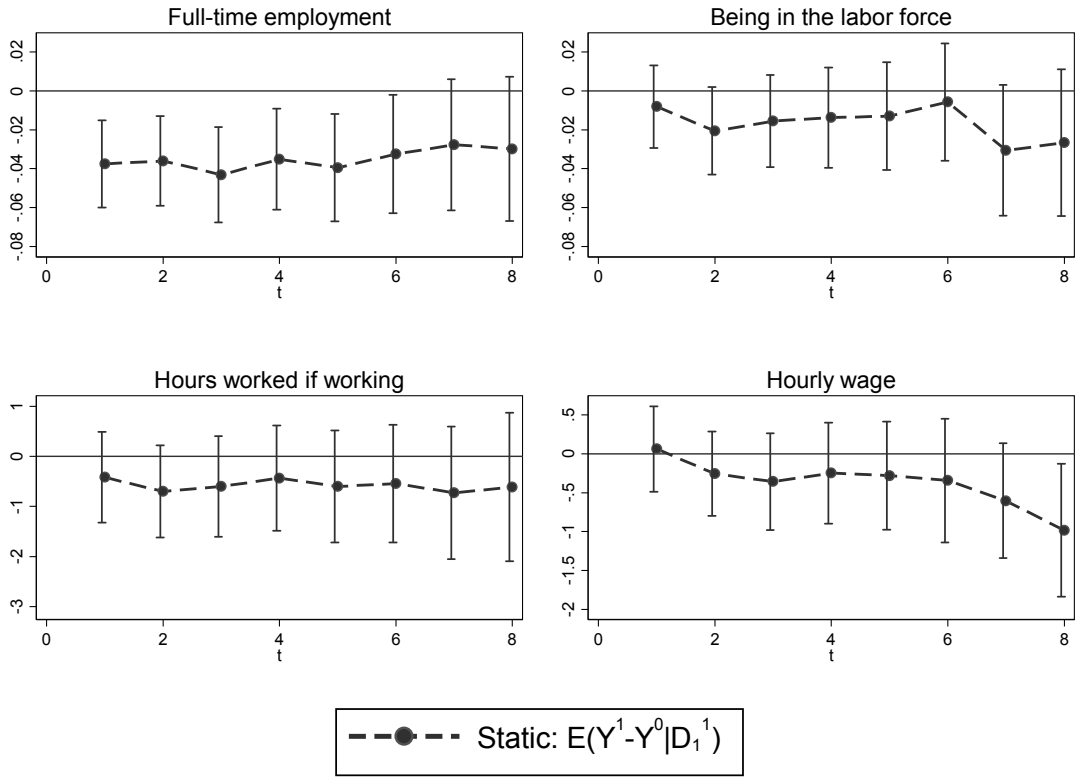


Figure 3.4: Labor market effects of informal caregiving for females – Static version

Source: SOEP. Own calculations. Note: The graph shows the point estimates and the 95% confidence intervals.

The estimates on being in the labor force (upper right panel) are very close to zero and thus insignificant throughout. Conditional working hours (lower left panel) are reduced by a little less than one hour, on average. Over time, the effect is persistent and slightly increasing but never significant and always rather small in magnitude. A quick back-on-the-envelope calculation shows how these findings fit together. Given that 76% of women in the sample are employed and 35% work full-time (42.2 hours on average, actual working time), 41% work part-time (22.2 hours on average). Assuming that no women leaves the labor force due to care provision but 4%-points switch to part-time work reduces the average conditional working hours from 31.4 to 30.4. This would imply a reduction by one hour, not far from the observed effect.

Another potential labor market effect are wage penalties. This is explored in the lower right panel of Figure 3.4 where we assess the effect on hourly wages for all employed females. This graph reveals zero contemporaneous wage-effects. It seems to be, however, that fairly small negative effects add-up over time and become sizeable and significant after a couple of years. One reason for this could be forgone promotions due to caregiving with lagged effects that only materialize some time after care provision. Eight years after care provision, working women have a by 1 Euro per hour lower gross hourly wage which is around 7% in relative terms.

While there are, to our knowledge, no studies on informal care that evaluate the wage effects over a similar time interval, there are, however, related papers on more general forms of dynamic labor market repercussions, e.g., due to health shocks or fertility decisions that are to some extent comparable to our findings. Thus, these papers can provide a framework for why we do not observe short- but rather long-term wage effects. For example, García-Gómez et al. (2013) provide evidence for an increasing effect of a health shock on personal income overtime. The authors partly attribute this effect to a modest dynamic effect on labor market earnings (direct hourly wage effects are not reported). In contrast to this, Adda et al. (2017) employ a structural dynamic model in a recent paper and show that the career cost of children accumulate over time by "a combination of occupational choice, lost earnings due to intermittency, lost investment into skills and atrophy of skills while out of work, and a reduction in work hours when in work." This mechanism may transfer similarly to our context of informal care, providing a suggestive explanation for why the wage effect that we identify magnifies over time.

How do the short-run effects compare to those found in the previous literature that typically uses instrumental variables approaches? Answering this question is not simple as the range of estimates found in these studies is fairly large, see the introduction for citations of the studies. For instance, the range of caregiving effects on the probability to work goes from around 0 to a 30 percentage point decrease. Mixed results also hold for the effects on hours worked and wages.

Given that this study aims at estimating average treatment effects on the treated while studies using different instrumental variables probably identify different local average treatment effects it is hard to directly compare the findings.[17] What can be said, however, is that our short-run estimates are among the more conservative findings and in the lower – but not completely unusual – range of estimates from the previous literature. It should also be noted, however, that our approach is not too far away from the IV literature where several (also recent) studies do not reject the hypothesis of exogeneity of care provision for at least important parts of their specifications and fall back to OLS or fixed effects regressions. With respect to our most important long-run effect – small but persistent effects on full-time work and its interpretation as a switch to part-time work – our results are in line with Skira (2015) who finds that "women face low probabilities of [...] increasing work hours after a caregiving spell."

[17] Note, however, that the average treatment effect on the treated is also likely to differ from the average treatment effect.

3.5 Sensitivity analysis

So far we interpreted our estimates as causal conditional on the validity of our identifying assumption (the CIA) and found, in particular, considerable and persistent negative effects on full-time employment. We justified the identifying assumption by fully exploiting the panel information in the SOEP and using many observable characteristics to match on. As we also match on pre-treatment outcomes, time-constant unobserved heterogeneity that probably explains a lot of the willingness and ability to provide care is also taken into account.

Figure 3.5 reports results of a test whether there are potential anticipatory effects of future care provision. To do so, we repeat the baseline static matching procedure from Figure 3.4 also to outcomes in the years before treatment. More specifically, we include placebo estimates that, as an example, use care in $t = 1$ as the treatment, full-time work in $t = -2$ as an outcome and all other controls in $t = 0$ to match on.[18] Apparently, there is no significant pre-treatment change in full-time employment due to later care provision. The only significant drop here takes place in period one, which is the already familiar and persistent 4 ppts. reduction.

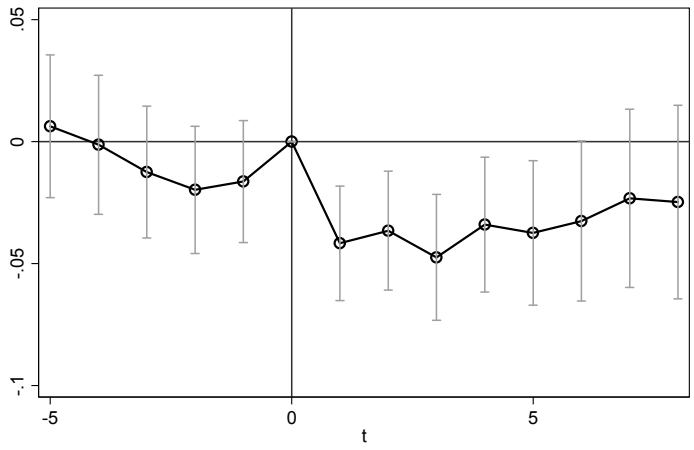


Figure 3.5: Pre-treatment trends for full-time employment

Source: SOEP. Own calculations. Note: The graph shows the point estimates and the 95% confidence intervals.

Nevertheless, there might be other confounding factors that we do not observe. Whether this is the case in our study is inherently non-testable. Thus, we do not and cannot say anything in this section concerning the likelihood that our assumption is fulfilled. Rather, we want to ask how crucial certain plausible deviations are for our results. As the CIA is not an "all or nothing" assumption, different degrees of its violation still allow to bound causal effects and to receive meaningful parameters. Other robustness checks, as outlined in Section 3.4.3, are presented in the Supplementary Materials.

[18] We restrict robustness checks and sensitivity analyses to full-time work for sake of brevity, as considerable effects are only found for this outcome.

3.5.1 General framework

In this section we follow an approach by Ichino et al. (2008) who refined the suggestions for sensitivity analyses by Rosenbaum and Rubin (1983) and Imbens (2003) and implemented them in a more practical and easy to interpret fashion. This analysis is also in the spirit of the one suggested by Altonji et al. (2005) without the need to make strong parametric assumptions.

We now assume that the CIA does not hold

$$Y_t^0 \not\perp\!\!\!\perp D_1 | X_0, Y_0 \qquad \forall t > 0$$

but that the failure is due to an unobserved variable U_0. Could we condition on it, we had

$$Y_t^0 \perp\!\!\!\perp D_1 | X_0, Y_0, U_0 \qquad \forall t > 0.$$

Hence, all the unobserved heterogeneity that leads to assumed endogeneity problems is captured by U_0. For simplicity, we assign U the time indicator 0 but it could also be a variable measured between 0 and 1. To keep things as simple as possible, Ichino et al. (2008) follow Rosenbaum and Rubin (1983) who proposed U_0 to be binary. This is appealing, since the distribution of a binary variable is fully determined by its mean. To describe how U_0 affects both treatment and outcome, four probabilities p_{ij}, $i \in \{0,1\}$; $j \in \{0,1\}$ are defined as

$$\begin{aligned} p_{01} &= Pr(U_0 = 1 | D_1 = 0, Y_t = 1) \\ p_{00} &= Pr(U_0 = 1 | D_1 = 0, Y_t = 0) \\ p_{11} &= Pr(U_0 = 1 | D_1 = 1, Y_t = 1) \\ p_{10} &= Pr(U_0 = 1 | D_1 = 1, Y_t = 0). \end{aligned} \tag{3.1}$$

Treatment status D_1 and binary outcome category Y_t are observed in the data and, hence, individuals can be assigned one of the four probabilities p_{ij} where i denotes treatment status and j the outcome. The four above equations fully define the distribution of the hypothetical confounding variable U_0. Differences in these probabilities will mechanically introduce a correlation between U_0 and both, D_1 and Y_t and, thus, U_0 will be an important confounding factor.

Given values for p_{ij} we simulate U_0 by drawing 200 times from Bernoulli distributions with the respective parameters for each individual and estimate the ATT 200 times, conditioning on X_0 and Y_0 as before, but also on U_0. Taking the average over all results provides us with point estimates as well as standard errors of the average treatment effect where the CIA is relaxed.[19]

We follow Ichino et al. (2008) and set p_{ij} such that we control the "outcome effect" (the relationship with Y_t) and the "selection effect" (the relationship with D_1) of U_0. As an illustration, think of U_0 as a health shock again that both affects the probability to work and to provide care. $U_0 = 1$ indicates a health shock, $U_0 = 0$ means no health

[19] We use a modified version of the Stata command sensatt that is written by Nannicini (2007).

shock. This unobserved variable certainly has a negative and strong selection effect such that unhealthy people are less likely to provide care. It may also have a negative outcome effect. More formally, Ichino et al. (2008) define the parameter $s = p_{1.} - p_{0.}$ as the selection effect where

$$p_{i.} = Pr(U_0 = 1|D_1 = i) = p_{i0} \cdot P(Y_t = 0|D_1 = i) + p_{i1} \cdot P(Y_t = 1|D_1 = i) \quad i \in \{0,1\}\,.$$

The larger this effect, the larger is the effect of U_0 on selection into treatment keeping the outcome fixed. The outcome effect, defined as $d = p_{01} - p_{00}$ reflects the correlation between U_0 and the untreated counterfactual outcome. As an example, an outcome effect of $d = -0.05$ means that, in the group of non-carers, among those who work the likelihood to experience a health shock is 5 percentage points smaller. The higher d the stronger is this correlation. Likewise, a selection effect of $s = -0.05$ implies that among caregivers the likelihood of $U_0 = 1$ is lower than among non-caregivers. Given these settings, U_0 is a variable that is both correlated with treatment and outcome and should be accounted for in the estimations. Once we set values for d and s we can derive the four p_{ij} by solving an equation system and simulate U_0.[20]

3.5.2 Choice of selection and outcome effects

In principle, d and s could be arbitrarily chosen and certainly such that the identified effect of care provision on full-time work turns zero or even positive. This, however, does not deliver any useful information as it is always possible to reduce the set of assumptions so far that zero is included in the identified bounds (see the worst-case bounds by Manski, 1995, that always include a treatment effect of zero). Thus, a major challenge is to find reasonable deviations from the CIA.

We follow the reasoning by Altonji et al. (2005) and argue that we have a high quality panel data set that allows to observe a large amount of variables determining care provision and labor force participation. Among them are baseline health, age, education, preferences, and pre-treatment outcomes. As noted above, we are able to explain 67% of the variation in full-time work by our observable factors. Thus, there is room for unobservables. However, their impact is probably not drastically larger than the impact of the observables.

Thus, one way to find reasonable values for d and s is to go back to equation system (1) and – starting the other way around – use observed binary variables in the data set, substitute them for the unobserved U_0 and calculate the selection and the outcome effect of these variables. Thus, we get a feeling how selection and outcome effect of important and observed variables are distributed in the data. Next, one could argue that the unobserved variable U_0 might have a similar selection and outcome effect as important observed variables. We follow this approach and compute these effects for all variables in the sample. Results are reported in Figure 3.6 and, more detailed, in Table 3.7 in the supplementary materials.

Figure 3.6 reveals that most of the variables have selection and outcome effects of at most 0.1 in absolute values. The dashed black line marks the interval of selection

[20] Using the two equations above, also assuming that $p_{11} - p_{10} = 0$ and assuming a value for $P(U_0$ we have four equations and can determine the four unknown p's.

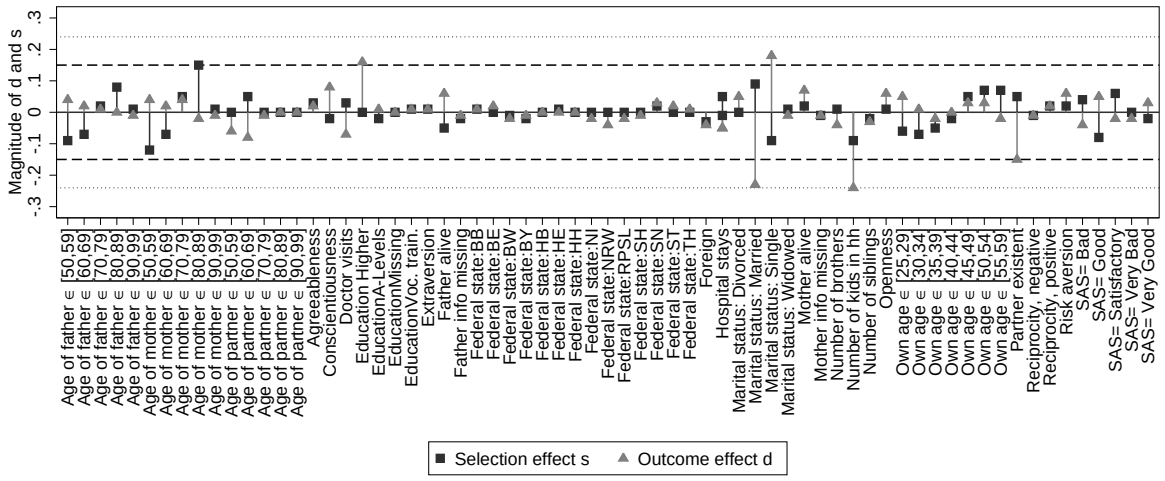


Figure 3.6: Parameters for calibration of the sensitivity analysis

Source: SOEP. Own calculations. Note: See Table 3.4 for translations of variable names. All values are reported in Table 3.7 in the supplementary material. Full-time work is not shown here. Year dummies are not reported for legibility but have values of d and s of virtually zero.

effects all observable variables fall in (± 0.15), while the grey dotted line (± 0.23) does the same for the outcome effect. The only exception is the pre-treatment outcome full-time work – which is left out in the figure – that has a selection effect of $s = -0.05$ (among the carers, the pre-treatment full-time employment rate is 5 ppts. lower), and an outcome effect of $d = 0.8$ (the probability to have worked full-time in $t = 0$ is 80 ppts. higher among those who work in $t = 1$ than those who do not work in $t = 1$). The huge value for full-time work mainly reflects the path dependence in full-time work and, therefore, this pre-treatment outcome variable is not very helpful to find credible values for s and d.

We test three different combinations of d and s to bound the treatment effects under weaker assumptions than the CIA. First, we calibrate a confounder U_0 to have the same bivariate correlation with treatment and outcome as pre-treatment full-time work. Even though this should not be the most interesting case as this variable is somewhat particular, this provides a first benchmark. Next, we assume a left out variable U_0 that has a correlation much stronger than all other observed variables by using $d = -0.24$ and $s = -0.15$. If there was indeed such a variable and we took this into account, the treatment effect should increase in absolute values (get further away from zero). The example for such a variable discussed before was a health shock or the death of a spouse. It should be noted, however, that even the health shock considered in Section 3.9 has outcome and selection effects of far less than 0.1 and, thus, should be exceeded by this variable. In total, this variable U_0 is linked much stronger to treatment and outcome than any of the observed variables.

As the second specification increases the treatment effect, we want to challenge our results by using $d = -0.24$ and $s = 0.15$. This parameter combination will push the treatment effects towards zero. As it seems to be hard to imagine a variable with such a drastic effect and in a direction opposed to the one discussed by a health shock, this

could be seen as a credible lower bound of the true effects. Again, note that a handful of variables either has a larger d or s but none has both parameters at such high levels.

3.5.3 Results

Figure 3.7(a) shows the resulting effects for a model that includes U_0 calibrated to have the same effect as the pre-treatment outcome (light grey triangles) and relates it to the same model without such a confounding factor (represented by black circles; shown before in the upper right panel of Figure 3.4). The difference in the effects of both models is statistically indistinguishable and also the magnitude of the effect is fairly similar.

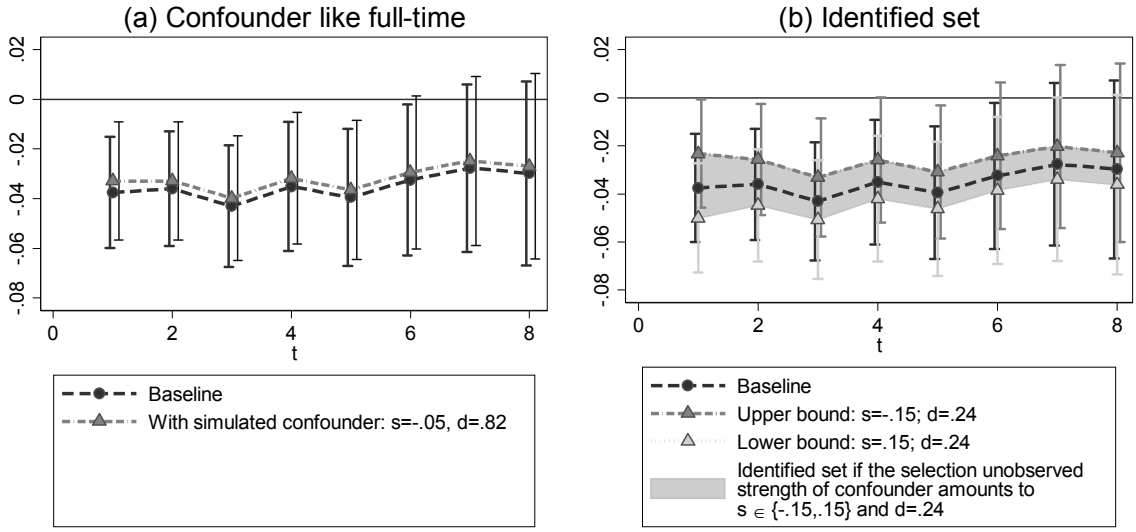


Figure 3.7: Sensitivity analysis for full-time employment

Source: SOEP. Own calculations. Note: The graph shows the point estimates and the 95% confidence intervals.

Figure 3.7(b) shows the set that we can identify if we had knowledge that the impact of the unobserved heterogeneity is bounded between $s \in \{-0.15, 0.15\}$ and $d = 0.24$ (grey-shaded area). The resulting estimates are bounded between -2.3 and -5.0 percentage points. If there was a variable we left out that worked like a health or other life event shock but had a much stronger impact than observed health shocks by our definition above, the true effect would be around -5.0. If the left-out variable had an opposing effect on treatment or outcome and, again, it had a very strong effect on both, we would still identify an effect of -2.3 percentage points.

To sum up, even if there was another confounding factor with effects as extreme as the pre-treatment outcome or considerably stronger than any of the other observed controls: conditioning on it would only partly reduce the magnitude of effects. Given the large amount of other variables we control for – and that these variables have much smaller selection and outcome effects even though these are variables as important as age, education, parental characteristics and personality traits – it seems hard to imagine unobserved variables with even more drastic effects that would, if we conditioned on them, destroy the results. Thus, even if the CIA were not to hold,

our effects would remain fairly stable and all our conclusions from Section 3.4 sustain.

3.6 A dynamic design

3.6.1 Empirical strategy II

The previous approach answers a relevant question: given that a women provides care today, what effects can she expect for her labor force status today, in one year, in eight years? Given that the treatment is defined in year 1 only, this effect is a mixture of different care provision paths later on. The treatment group consists of individuals that provide only one year of care but also of those who care for two consecutive years, three years and any other care spell (like care, no care, care, no care,....). Likewise, the control group includes individuals who take up care provision later on.[21]

A potentially more interesting ideal experiment would be to assign women randomly to different paths of care. By this means, not only the start of caregiving is randomized, as in the static setting. Also, the selection out of care is controlled for by dynamic attributes. The main advantage of such an approach is that we can relate effects of different care paths to one another, for instance, in order to see whether the static effect is dominated by one particular path. Thus, it seems natural to ask whether the (long-run) effect of providing more consecutive years of care differs from providing (at least) one year. This, however, considerably complicates estimation as time-varying control variables and outcomes along the care provision path potentially affect the decision to stay caregiver or to cease. Lechner (2009b) suggests an approach that is able to capture the effects of different treatment paths where it is decided sequentially at different nodes on a decision tree whether the treatment is continued or stopped (or, more generally, another treatment is taken). In an example, Figure 3.8 shows an excerpt of potential paths (D_1, D_2, D_3), where $D_t \in \{0, 1\}$ for $t = 1, 2, 3$, that can be taken on in three years. Dynamic matching/reweighting means that we add a time dimension to the matching/reweighting process. As static matching aims to control for any differences in observed characteristics just prior to the caregiving decision, dynamic matching also balances time differences in the controls that may influence any particular care path.

We argue that this partly dynamic modelling of the care dynamics – up to three consecutive years – is sufficiently interesting, and that we do not need a full dynamic model over the 8-year period to learn about the most important average impacts of care provision. Importantly, most careprovision spells in our data set have a short term nature. 60% of care spells in the sample last for one period, 17% for two periods, and 7% for three (see Section 3.3). Yet, it is potentially relevant to model the dynamic decision for three years (instead of fully sticking to the static model) at least for the following reason. Often, when individuals enter their care spell, they do not

[21] However, see a robustness check in Figure 3.13 in the Appendix where we restrict the control group to individuals that never provide care throughout the full observation period. The results do not differ. This is not the preferred specification as the definition of the control group depends on future caregiver status.

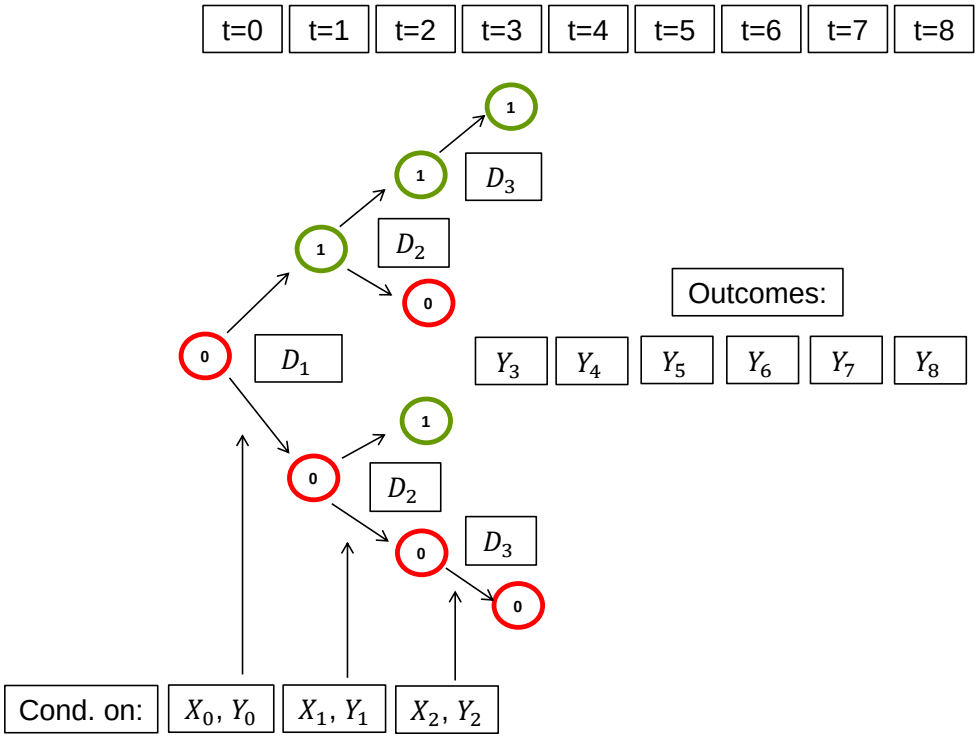


Figure 3.8: Dynamic design

Own illustration.

have enough information to build an expectation on the duration of the spell and on the burden they take on. Thus, it is conceivable that many individuals keep their labor market participation unchanged in the first period. It is probably fair to assume that those who have already provided care for two years have a fairly good (though certainly not perfect) idea of their ability to take on the double burden of working and caring and at least more information (if only vague) on the potential future duration of the care spell. Thus, while at the first node, many individuals potentially make an explicit short-term decision on their labor market participation during the care episode, this is more likely to be a longer term decision – meaning a decision for the full care episode – at the second or third node and it appears to be interesting to explicitly look at the effects of caring at least three consecutive years on labor market outcomes.

In the following we fully draw on Lechner and Miquel (2010) and only very briefly sketch their ideas for the dynamic model, restricting the outline to two periods. The interested reader is referred to Lechner (2008, 2009b) and Lechner and Miquel (2010) to find derivations and in-depths discussions of the full model (including identification and estimation). The observational rule for two periods of potential care provision reads

$$Y_t = D_1 D_2 Y_t^{11} + (1 - D_1) D_2 Y_t^{01} + D_1 (1 - D_2) Y_t^{10} + (1 - D_1)(1 - D_2) Y_t^{00}, \qquad t \geq 2$$

where D_2 is careprovision in period 2 and Y^{11} the potential outcome of caring two consecutive periods and analogously for the three other potential outcomes.[22]

We could be interested in the effect of caring for two consecutive years as opposed to not caring two consecutive years (for the group of individuals who provide care in the first year). This is a kind of an average treatment effect on the treated (for those treated in period 1) and can directly be compared to the ATT from the static version. In technical terms, we want to know

$$DATT_t = E(Y_t^{11} - Y_t^{00} | D_1 = 1)$$

where $DATT_t$ is called the dynamic treatment effect on the treated. Of course, effects for other differences in potential outcomes (that is, other treatment paths) and other subpopulations (e.g. the full population of individuals who did not provide care in $t = 0$) can, in principle, be calculated as well.

Estimation of this effect, again, amounts to finding observable outcomes that can be used to estimate the unobservable counterfactual outcomes. In essence, this is finding individuals that took on exactly the two paths ($D_1 = 1$, $D_2 = 1$) and ($D_1 = 0$, $D_2 = 0$) but share – except for the treatment, or parts of the treatment – the same characteristics as the subpopulation we want to calculate the $DATT$ for, here, the caregivers in period one, $D_1 = 1$. This amounts to the "weak dynamic conditional independence assumption" (Lechner and Miquel, 2010):

1. $Y_2^{00}, Y_2^{10}, Y_2^{01}, Y_2^{11} \perp\!\!\!\perp D_1 | X_0, Y_0$
2. $Y_2^{00}, Y_2^{10}, Y_2^{01}, Y_2^{11} \perp\!\!\!\perp D_2 | X_1, X_0, Y_1, Y_0, D_1$

This means that conditional independence is assumed to hold at each node and is achieved by sequentially modelling all transitions between two years (e.g., the one from $t = 0$ to $t = 1$, then the one from $t = 1$ to $t = 2$ and so on) and, thereby, conditioning on each node for the full set of pre-treatment control variables. For instance, at the transition from $t = 1$ to $t = 2$ we control for X_0 and X_1 and, again, previous outcomes Y_0 and Y_1. This explicitly allows for individuals who started to provide care in $t = 1$ and then, in $t = 2$, stopped caregiving due to effects of care provision on either control variables (for instance a drop in own health) or on labor market outcomes. Given that characteristics of potential care recipients are also in the set of controls, this also allows for stopped care provision because there was no need anymore (e.g., because the care recipient passed away). Thus, we explicitly take into account changes in control variables and outcomes over time to explain why individuals take on different treatment paths. The remaining reasons to choose different paths are assumed to not be systematically related to the individual's potential outcomes.

Estimation, thus, involves several steps that we outline here. In contrast to the static version we fully restrict the analysis to inverse probability weighting and do not use matching.[23]

[22] Note, when $t > 2$, $Y_t^{k,l}$, $k, l \in \{0, 1\}$ is a mixture of all those potential outcomes that follow in the care path after the sequence $[k, l]$.

[23] This is mainly to fully follow Lechner (2009b) who only uses IPW for the dynamic model.

1. Estimate the propensity score for the decision on the first node ($Pr(D_1 = 1|X_0, Y_0)$) and the two propensity scores (depending on the decision in the last period) for the second node ($Pr(D_2 = 0|D_1 = 0, X_1, X_0, Y_1, Y_0), Pr(D_2 = 1|D_1 = 1, X_1, X_0, Y_1, Y_0)$)

2. Define the relevant dynamic treatment and control group

$$D = \begin{cases} 1 & \text{if } (D_1 = 1) \cdot (D_2 = 1) = 1 \\ 0 & \text{if } (D_1 = 0) \cdot (D_2 = 0) = 1 \end{cases}$$

3. Compute the inverse probability weights:

$$W = \begin{cases} \dfrac{1}{Pr(D_2 = 1|D_1 = 1, X_1, X_0, Y_1, Y_0) \cdot Pr(D_1 = 1|X_0, Y_0)} & \text{if } D = 1 \\ \dfrac{1}{Pr(D_2 = 0|D_1 = 0, X_1, X_0, Y_1, Y_0) \cdot (1 - Pr(D_1 = 1|X_0, Y_0))} & \text{if } D = 0 \end{cases}$$

4. In order to make our estimator less sensitive towards very high or very low propensity scores we only keep observations within the 5$^{\text{th}}$ and 95$^{\text{th}}$ percentile of the $Pr(D_1 = 1|X_0, Y_0)$ distribution. Furthermore, we condition on the common support of $Pr(D_1 = 1|X_0, Y_0)$, $Pr(D_2 = 1|D_1 = 1, X_1, X_0, Y_1, Y_0)$, and $Pr(D_2 = 1|D_1 = 1, X_1, X_0, Y_1, Y_0)$ respectively.[24]

5. Then the dynamic average treatment effect amounts to:
$DATT_t = (D'WD)^{-1}D'WY_t$

3.6.2 Estimation results – Dynamic model

Figure 3.9 adds to the static effects the estimated effects of providing care in both year 1 and 2 compared to not providing care in both years ($E(Y_t^{11} - Y_t^{00}|D_1 = 1)$) as well as providing care in all first three years compared to not caring then ($E(Y_t^{111} - Y_t^{000}|D_1 = 1)$). A first general and main result is that, for full-time employment and conditional hours worked, it does not make a difference whether one looks at the effect of caregiving for at least one year or to caregiving for at least two or three consecutive years. Most point estimates do not differ significantly. While this is partly due to larger standard errors in the dynamic estimations, the point estimates are mostly also quite similar in magnitude, too. Individuals who have been providing care for three periods do not have a significantly lower probability to work-full time due to care provision in the third year than those who provided care for one year – the effect is a 5 ppt. reduction compared to 4 ppts. for one year.

Thus, contrary to what we expected, individuals seem to directly decide in the first year of care provision about their short- and medium-run labor market participation. Or, put differently, there seem to be no dynamic effects – at least for the period of three years – of care provision on the likelihood to work full-time in Germany.

[24] See Figure 3.21 in the supplementary materials for an exemplary visual overview of the propensity score distributions as well as the exact number of observations that are dropped for every single restriction.

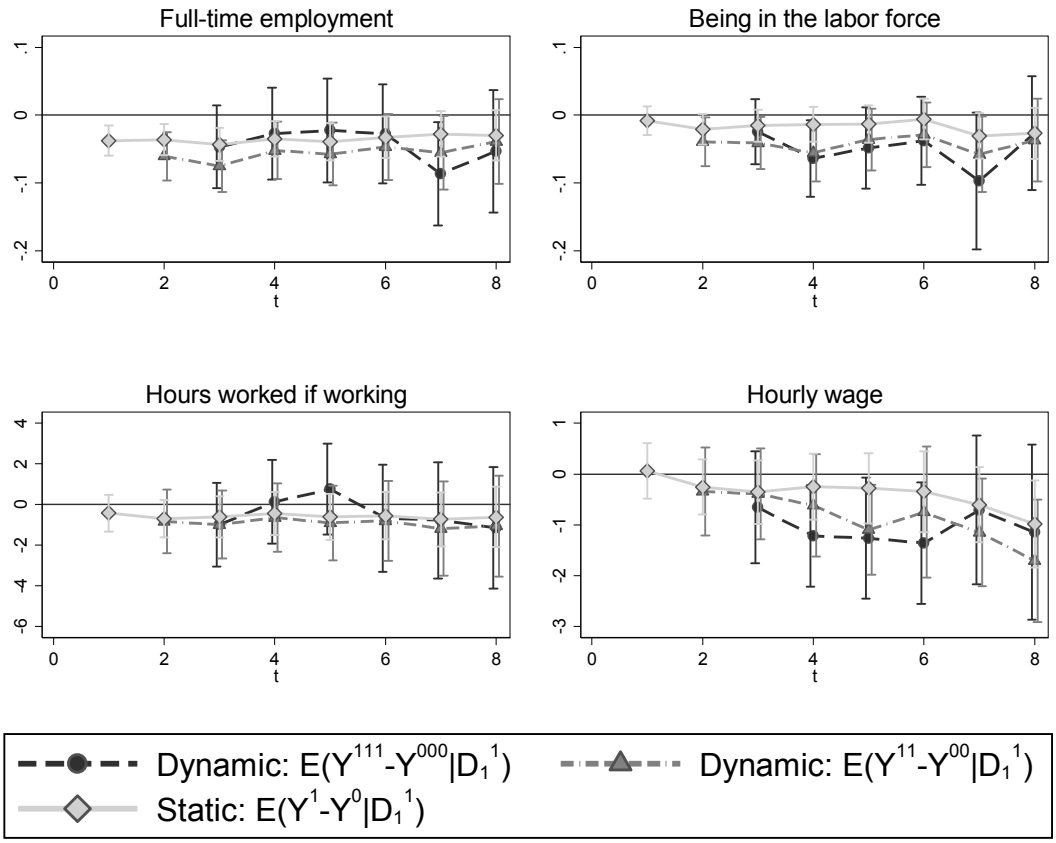


Figure 3.9: Labor market effects of informal caregiving for females – Dynamic version

Source: SOEP. Own calculations. Note: The graph shows the point estimates and the 95% confidence intervals.

Moreover, this picture does not change in the long-run perspective until several years after care provision. For full-time work and conditional hours, the dynamic effects are somewhat more pronounced, but the differences are rather small.

The findings are somewhat different for the probability to be in the labor force and for wages where we find differences between different care paths. Here it seems that longer care spells translate into higher effects. As opposed to the static model, the probability of being in the labor force is reduced by around 3–6 ppts. for longer-lasting care spells. Except for the fourth year, these effects are not statistically significant, however. For wages, caring at least three consecutive years goes along with a significant wage penalty of nearly 2€/hour (around 14% in relative terms). However, considerably smaller sample sizes and less degrees of freedom for the dynamic specifications also add more noise to the results. Given the general comparability of both approaches, we do not repeat the robustness checks of Section 3.5 here, and, for alternative specifications that are also data demanding, turn back to the static version.

3.7 Alternative specifications of the treatment variable

An important question is how sensitive the effects are with respect to the care intensity. In the baseline specifications, we defined the treatment to be at least one hour of care per day. In the following we vary this definition by restricting the treatment to at least two hours or three hours per day. The number of observations in the treatment group is then reduced from 2,186 to 847 for two hours of care per day and to 380 for three hours. Figure 3.10 – which returns to the static version due to sample size reasons as well as comparably small differences between static and dynamic approach – compares the results for these definitions with the baseline results. Apparently, there are hardly any differences between one and two hours of care per day, both in the short- and the longer-run. Moreover, short-run effects (effects in $t = 1$) also do not differ between three daily hours and one hour as a treatment definition for any of the four outcome variables. However, longer run effects for full-time work and employment are stronger if we use the cut-off of three hours. This is remarkable as the strongest effects seem to materialize when, in most cases, the care episode has already ceased. Those who provided care of at least three hours per day are, 8 years later, around 15 ppts. less likely to work full-time and to be employed. Moreover, those who stay in the labor force earn, on average, 2 Euro per hour less, which is a considerable wage penalty.

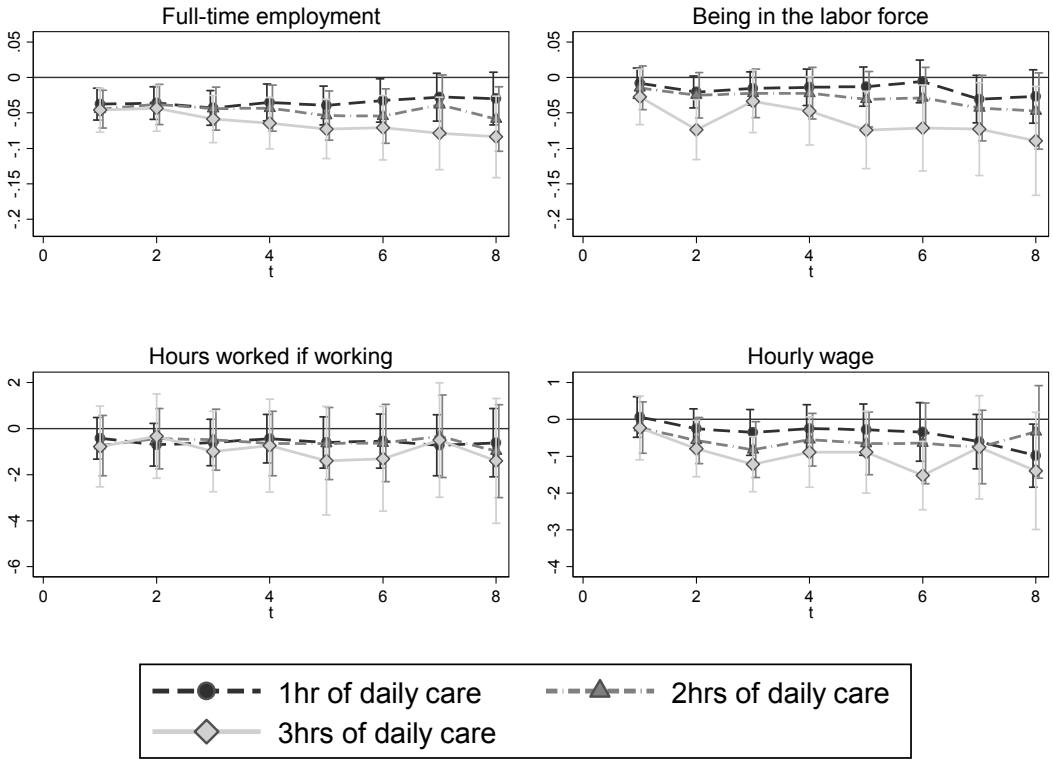


Figure 3.10: Results of the static version – Variations in treatment definition

Source: SOEP. Own calculations. Note: The graph shows the point estimates and the 95% confidence intervals.

One issue in estimating longer run effects is that, due to our sample construction, the average age of women in $t = 8$ is 48 and, thus, higher than in $t = 1$ (44). As long as

effects of care provision are not heterogeneous in age, this should not be a problem. Nevertheless, the results might be affected by women who anyway leave the labor force when they get older. As an example, a 60 year old women in $t = 1$ automatically drops out of the sample in $t = 6$ and longer-run effects can only be estimated for women who are at most 57 years old when they start to provide care. Figure 3.14 in the Appendix repeats the analysis but restricts women to be up to 55 years in $t = 1$. The results are largely unchanged for the whole set of outcome variables. Thus, our estimates do not seem to be affected by either the mandatory retirement threshold nor other effects related to aging. We provide more analyses on age differences in the supplementary materials where we split the sample at median age of 44 and look at differences between younger and older women. We find that differences are quite small.

3.8 Conclusion

In this paper we assessed labor market outcomes as an important part of the implicit costs of informal care provision. In order to identify these costs we use matching techniques and inverse probability weighting. We exploit the panel information and a large set of individual controls (including measures of personality traits) to justify the identifying assumptions but also relax the assumptions in sensitivity analyses. We compare effects of providing care in a certain year on contemporaneous and later outcomes to effects of up to three consecutive years of care provision. Thereby, we contribute to the literature by both analyzing longer run effects as well as explicitly taking into account the dynamics of care provision.

An overview of our results is given in Table 3.3. Most importantly, we find significant initial negative effects of informal care provision on the probability to work full-time. The 4 percentage points reduction in the probability to work full-time after caring for at least one year is persistent over time. These effects are largely comparable for women who provide care for at least three consecutive years. Providing care for a higher intensity (at least three hours per day) has a stronger long-term effect on full-time work. Conditional working hours are, on average, only slightly affected. Long-run effects are reductions around 1 hour per week, which are also statistically insignificant. There are no short-run effects on the likelihood of being in the labor force but quite considerable negative effects for both longer care episodes and higher care intensities. Hourly wages are not affected in the short-run but we find a long-run wage penalty of around 1 to 1.5 Euro for women who provide care (irrespective of duration and intensity). Alternative specifications show that the effects are not only driven by older women who provide care.

We scrutinize our results by versatile tests to check whether they still hold even if there are deviations from our identifying assumption. For example, by simulating an additional confounder with a selection effect stronger than all observed ones, we are able to credibly bound the effect on full-time between 2.4 and 5.0 percentage points but argue, that if any, 5.0 should be more likely than 2.4.

The reduction in full-time work seems to be mostly driven by the intensive margin of labor supply. Women do not leave the labor market – at least for shorter durations

Table 3.3: Summary of results

Outcome	Care episode	Care provision / day	Short-run effect	Long-run effect
Full-time employment	$\geq$ 1 year	$\geq$ 1 hour	$-$ 4 ppts*	$-$ 4 ppts*
	$\geq$ 3 years	$\geq$ 1 hour	$-$ 4 ppts*	$-$ 4 ppts*
	$\geq$ 1 year	$\geq$ 3 hours	$-$ 5 ppts*	-9 to $-$ 15 ppts*
Conditional working hours	$\geq$ 1 year	$\geq$ 1 hour	≈ 0	-1
	$\geq$ 3 years	$\geq$ 1 hour	≈ 0	-1
	$\geq$ 1 year	$\geq$ 3 hours	-1	-2
Being in the labor force	$\geq$ 1 year	$\geq$ 1 hour	≈ 0	≈ 0
	$\geq$ 3 years	$\geq$ 1 hour	≈ 0	$-$ 3 to $-$ 6 ppts
	$\geq$ 1 year	$\geq$ 3 hours	≈ 0	-8 to $-$ 15 ppts*
Hourly wages	$\geq$ 1 year	$\geq$ 1 hour	≈ 0	€ -1*
	$\geq$ 3 years	$\geq$ 1 hour	€ -1.5*	€ -2*
	$\geq$ 1 year	$\geq$ 3 hours	≈ 0	€ -2*

Source: SOEP, own calculations. Summary of the results of different specifications as reported in Sections 3.4, 3.6, and 3.7. Short-run effect is one year after the start of a care spell (or after three years for care episodes of at least three years). Long-run effect is 7 - 8 years after start of a care spell. * indicates significance at the 5% level.

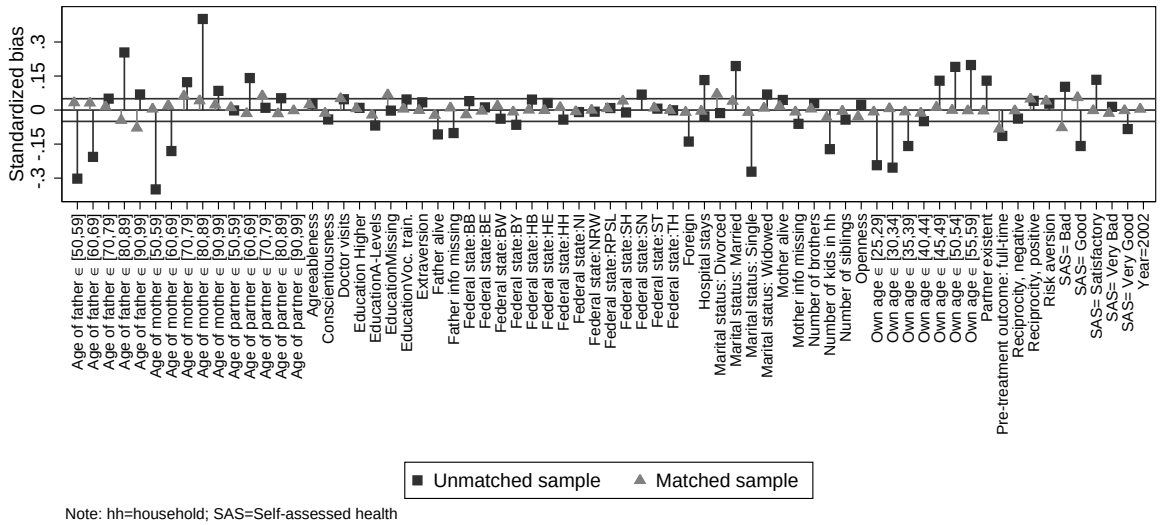
and moderate care intensities – but switch to part-time work. Yet, after the care spell has ceased, these women do not seem to switch back to full-time work. From a social planner's point of view, these effects would directly translate into costs and would weaken at least one argument in favor of informal care as opposed to other modes of care (informal care is usually assumed to be cheaper for the society). The following back-of-the-envelope calculation may elucidate this argument by showing that the estimated labor market responses due to informal care go along with fiscal costs. See Table 3.8 in the supplementary materials for details on the following derivations. Over the time span of eight years, on average, the females reduced their hours worked from 32.24 to 31.66 on the intensive margin. Together with the reduced employment probability (reduction from 77% to 75% on average over eight years) and the average wage penalty of female caregivers, caregiver's total labor market income would decrease from 18,224.06€ to 17,037.48€. For these incomes, income taxes amount to 2,229€ or 1,922€. Assuming a constant average consumption rate for caregivers and non-caregivers, one can calculate a resulting differential in the absolute amount of paid value-added-tax[25] which amounts to 123.33€. In total, according to this simplified calculation and based on our estimates, informal caregivers pay 430.33€ less taxes each year. For 2 million female caregivers currently in Germany, the resulting total tax differential due to informal care is estimated to be 860.66 million € per year. Note, however, that these numbers are largely based on insignificant estimates and should not be taken as granted. Moreover, we do not claim this to be the most important effect as, e.g., a loss of hourly wages by only one dollar after a couple of years roughly translates into 1,800 Euro (gross) per year for a full-time position (=40 hours $\times$ 45 weeks) and has cumulative effects for pensions later on, too.

[25]The value added tax is calculated as: $(18,228.06 - 2,229) - (17,037.84 - 1,922)$€ times the average value-added tax rate (weighted average between 19 and 7%, assumed to be 15%).

To the extent that the found effects are due to labor market frictions – something we did not show – and in case the government follows the goal of increasing female labor-market participation, the following policy measure could be thought of. A potential way to keep women in full-time work in the long run could be the expansion of the system of parental leave benefits to the informal care sector. Currently, German parents can leave their job for up to 14 months to care for their children and receive 60 per cent of their income (up to €1800). Expanding this to informal caregivers could fulfill two goals. First, caregivers could take a one year leave and do not need to take on the double burden of care provision and full-time work which probably has negative health effects (Schmitz and Stroka, 2013). Second, women are prevented from switching to part-time jobs to circumvent the double burden – apparently once women switched to part-time work they often do not switch back later. As long as caregivers have a legal claim to return to their previous job after one year (as is the case with the parental leave system) chances would probably be improved that informal care provision only has short-run but no long-run labor market consequences. Yet, even this would not prevent women from potential long-run wage penalties. Moreover, there are other potential options, including better coverage of formal ambulatory long care by the social long-term care insurance.

3.9 Appendix

Additional figures



Note: hh=household; SAS=Self-assessed health

Figure 3.11: Matching quality for full-time work

Source: SOEP. Own calculations. Note: The figure shows the normalized differences for both the unmatched and the matched sample. The normalized difference between treatment group (1) and control group (0) is calculated according to: $Diff = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{\frac{1}{2}(\sigma_1^2 + \sigma_0^2)}}$ where $\bar{x}$ is the sample mean and σ^2 the sample variance. Here, we report differences for all variables that are potentially included. Recall that only the subgroup of variables chosen by the double selection procedure (Belloni et al., 2014) is used in the propensity score estimations. See Table 3.4 for translations of variable names. An Epanechnikov kernel with of bandwidth 0.0018 is used. The two red lines mark a standardized bias of $\pm 5\%$. While a couple of variables falls outside this range in the unmatched sample this is only marginally the case for two of the variables in the matched sample. Pre-treatment full-time work as the most important control is highlighted in the figure. Year dummies are not reported for legibility but are well within the red lines and included in the estimations.

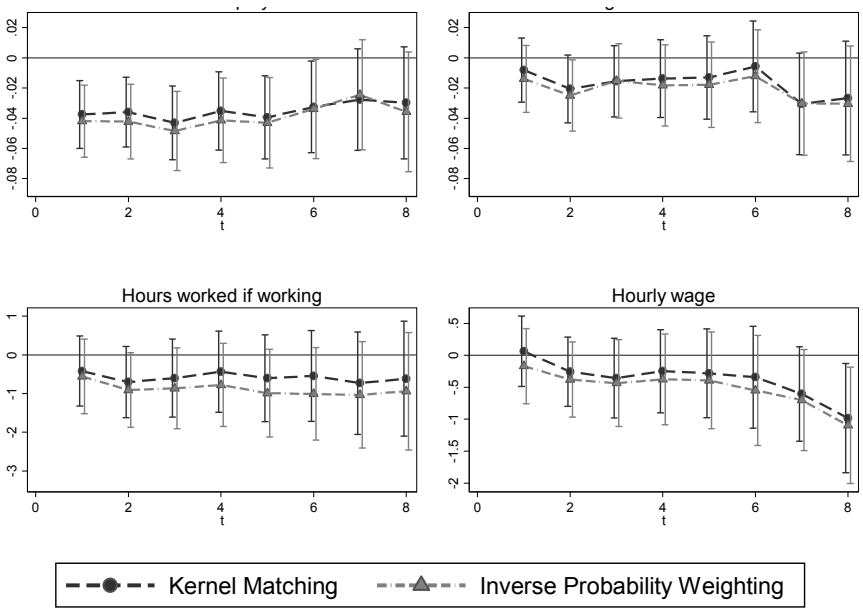


Figure 3.12: Static version, Kernel matching vs. IPW estimators

Source: SOEP. Own calculations. Note: The graph shows the point estimates and the 95% confidence intervals.

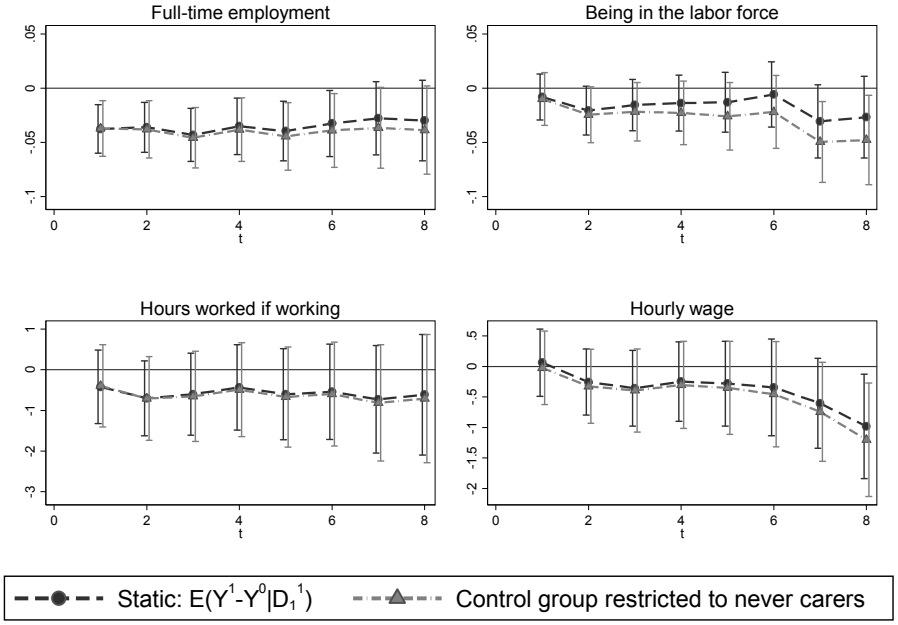


Figure 3.13: Difference between baseline results and same estimation with restriction to never carers in the control group

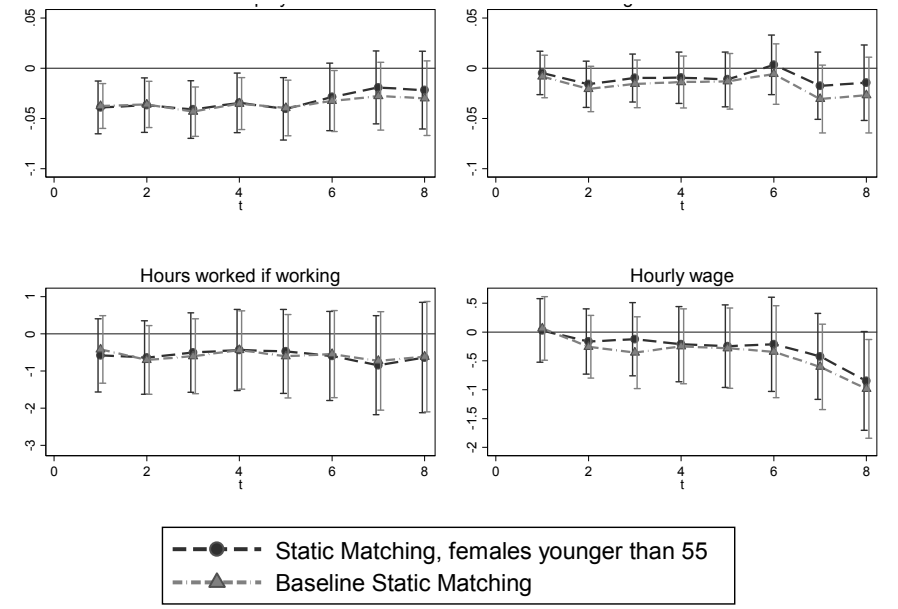


Figure 3.14: Results for females younger than 55 in $t = 1$

Source: SOEP. Own calculations. Note: The graph shows the point estimates and the 95% confidence intervals.

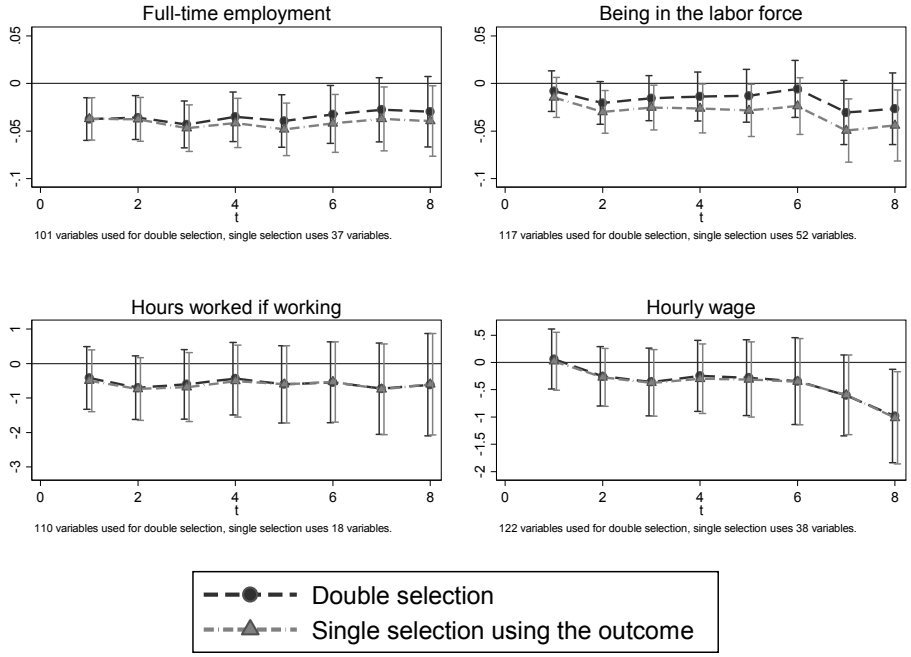


Figure 3.15: Double versus single post-lasso

Source: SOEP. Own calculations. Note: The graph shows the point estimates of the baseline specification (black circles) where the variables are selected using the double post lasso procedure. The grey triangles refer to the point estimates of the same effect with the sole difference that here the covariates are selected according to the single post lasso procedure. Consult the explanation in Section 3.4.2 for details. For both effects the 95% confidence intervals are reported. Regarding the difference in the results between both procedures, the effects are modestly magnified for two outcomes (full-time employment and being in the labor force) and fairly similar for the other two (conditional hours worked and hourly wage). In total, the effects seem to be not very sensitive. We therefore opt for double selection procedure with the more conservative results.

Additional tables

Table 3.4: Variable description

Variable	Description	Mean	SD	Min	Max
Outcome variables:					
Full time	Binary indicator of working full-time	0.35	0.48	0	1
Hourly wage	Gross monthly wage/(number of hours worked * 4.3)	14.04	8.75	2.0	149.5
Hours	Number of actual hours worked per week (here: unconditional)	21.89	18.69	0	80
Employed	Binary indicator, working full-time, part-time, vocational training, or marginal and irregular part-time employment	0.76	0.43	0	1
Care obligations:					
Age of mother					
≤ 59		0.20	0.40	0	1
$\in [60, 69]$		0.22	0.41	0	1
$\in [70, 79]$		0.20	0.40	0	1
$\in [80, 89]$		0.10	0.30	0	1
$\geq 90, 99$		0.02	0.13	0	1
Mother alive		0.73	0.44	0	1
Age of father					
≤ 59		0.14	0.34	0	1
$\in [60, 69]$		0.19	0.39	0	1
$\in [70, 79]$		0.16	0.37	0	1
$\in [80, 89]$		0.07	0.25	0	1
$\in [90, 99]$		0.02	0.14	0	1
Father alive		0.75	0.49	0	1
Partner existent		0.27	0.44	0	1
Age of partner		34.91	23.27	0	89
Number of siblings		1.86	1.73	0	18
Socio-economics and willingness to provide care:					
Own age					
$\in [25, 29]$		0.10	0.30	0	1
$\in [30, 34]$		0.12	0.32	0	1
$\in [35, 39]$		0.14	0.34	0	1
$\in [40, 44]$		0.16	0.36	0	1
$\in [45, 49]$		0.15	0.36	0	1
$\in [50, 54]$		0.13	0.34	0	1
$\in [55, 59]$		0.12	0.32	0	1
Education:					
A-Levels	Completed academic track	0.08	0.27	0	1
Voc. train.	Higher education and vocational training	0.07	0.25	0	1
Higher	Higher education	0.22	0.41	0	1
Missing	missing	0.01	0.12	0	1
Marital status:					
Single		0.17	0.38	0	1
Married		0.66	0.47	0	1

Continued on next page

Table 3.4 – continued

Variable	Description	Mean	SD	Min	Max
Divorced		0.10	0.30	0	1
Widowed		0.03	0.17	0	1
Foreign		0.08	0.27	0	1
Kids	Number of kids in the household	0.90	1.06	0	12
BIG-5 Inventory					
Neuroticism	Average of answers on 7-point scales	4.35	0.85	1	7
Conscientious-ness	Average of answers on 7-point scales	4.75	0.61	1	7
Agreeableness	Average of answers on 7-point scales	4.74	0.62	1	7
Openness	Average of answers on 7-point scales	4.59	1.21	1	7
Extraversion	Average of answers on 7-point scales	4.96	0.76	1	7
Reciprocity					
positive	Average of answers on 7-point scales	5.55	1.14	1	7
negative	Average of answers on 7-point scales	2.92	1.39	1	7
Risk aversion	Self-stated measure between 0 (very risk averse) and 10 (risk willing)	4.29	2.18	0	10
Ability to provide care:					
Self-assessed health (SAS)					
– Very Good	Binary: SAS = very good	0.09	0.29	0	1
– Good	Binary: SAS = good	0.45	0.50	0	1
–Satisfactory	Binary: SAS =satisfactory	0.32	0.47	0	1
– Bad	Binary: SAS = bad	0.12	0.33	0	1
– Very bad	Binary: SAS =very bad	0.02	0.16	0	1
Doctor visits	Number of doctor visits previous 3 months	2.56	3.85	0	99
Hospital stays	Number of hospital stays previous year	0.15	0.53	0	48
Year and Federal state dummies:					
Year					
=2002		0.10	0.30	0	1
=2003		0.09	0.29	0	1
=2004		0.09	0.29	0	1
=2005		0.08	0.28	0	1
=2006		0.09	0.28	0	1
=2007		0.08	0.27	0	1
=2008		0.08	0.26	0	1
=2009		0.08	0.27	0	1
=2010		0.07	0.25	0	1
=2011		0.08	0.27	0	1
=2012		0.08	0.27	0	1
Federal state:					
BE	Berlin	0.04	0.19	0	1
SH	Schleswig-Holstein	0.03	0.17	0	1
HH	Hamburg	0.01	0.12	0	1
NI	Lower Saxony	0.09	0.28	0	1
HB	Bremen	0.01	0.09	0	1
NRW	North-Rhine Westphalia	0.21	0.40	0	1
HE	Hesse	0.07	0.26	0	1

Continued on next page

Table 3.4 – continued

Variable	Description	Mean	SD	Min	Max
RPSL	Rhineland-Palatinate and Saarland	0.06	0.24	0	1
BW	Baden-Württemberg	0.12	0.33	0	1
BY	Bavaria	0.15	0.36	0	1
BB	Brandenburg	0.04	0.19	0	1
ST	Saxony-Anhalt	0.04	0.20	0	1
TH	Thuringia	0.04	0.20	0	1
SN	Saxony	0.07	0.25	0	1

Notes: Source SOEP

Table 3.5: Matching results corresponding to Figure 3.4

Outcome	Year	ATT	Std. err.	t-statistic	Observations
Full-time employment	1	−0.038	0.011	−3.282	65,307
	2	−0.036	0.012	−3.056	59,751
	3	−0.043	0.012	−3.448	52,383
	4	−0.035	0.013	−2.652	45,142
	5	−0.039	0.014	−2.806	38,303
	6	−0.032	0.016	−2.094	31,730
	7	−0.028	0.017	−1.608	25,820
	8	−0.03	0.019	−1.576	20,455
Conditional working hours	1	−0.42	0.462	−0.908	41,023
	2	−0.702	0.47	−1.494	37,018
	3	−0.601	0.514	−1.17	32,097
	4	−0.435	0.536	−0.812	27,446
	5	−0.602	0.571	−1.054	23,194
	6	−0.544	0.598	−0.909	19,189
	7	−0.728	0.675	−1.079	15,561
	8	−0.614	0.757	−0.812	12,266
Hourly wages	1	0.063	0.281	0.223	37,705
	2	−0.255	0.277	−0.923	34,092
	3	−0.357	0.317	−1.123	29,605
	4	−0.248	0.332	−0.747	25,368
	5	−0.281	0.354	−0.793	21,450
	6	−0.343	0.406	−0.844	17,785
	7	−0.604	0.377	−1.603	14,444
	8	−0.984	0.437	−2.252	11,420
Being in the labor force	1	−0.008	0.011	−0.746	65,294
	2	−0.021	0.011	−1.792	59,744
	3	−0.015	0.012	−1.286	52,378
	4	−0.014	0.013	−1.041	45,139
	5	−0.013	0.014	−0.918	38,299
	6	−0.006	0.015	−0.371	31,728
	7	−0.031	0.017	−1.781	25,823
	8	−0.027	0.019	−1.385	20,457

Source: SOEP, own calculations. Employed bandwidths: Full-time employment: 0.0017827; Conditional working hours: 0.0020926; Hourly wages: 0.0021498; Being in the labor force: 0.0018107.

Supplementary material

Additional results

Results to justify the CIA

Delete individuals with health shocks and loss of parents

Taking up the discussion of Section 3.4.3, two other reasons for potential failure of the CIA that come to mind are a health shock or a death of a parent or partner between period 0 and period 1. As, in particular, we cannot identify in the data whether a health shock between 0 and 1 was due to caregiving or, the other way around, care responsibilities were not taken up due to a health shock, we cannot account for a health shock, as this is potentially a "bad control".

In a robustness check, however, we identify all individuals who experienced a health shock or a death of a parent and exclude them from the analysis to see whether they affect the findings. Not uncommon to the health economic literature (see e.g., García-Gómez, 2011), we use the self-stated health on a 5-point scale to define a measure of health shock. In order to allow for a wide definition, we define a health shock as a deterioration to either "bad" (category 4) or "very bad (category 5)". This includes 4,537 person-year observations. A stricter condition of a reduction by at least two categories and to either "bad" (category 4) or "very bad (category 5)" is fulfilled by 1,643 person-year observations. Moreover, 1,005 had to suffer from the loss of a parent or spouse. Figure 3.16 reports the findings where individuals according to the wide definition of a health shock between 0 and 1 and those who have lost a parent are excluded from the sample. The results are statistically indistinguishable from the baseline results.

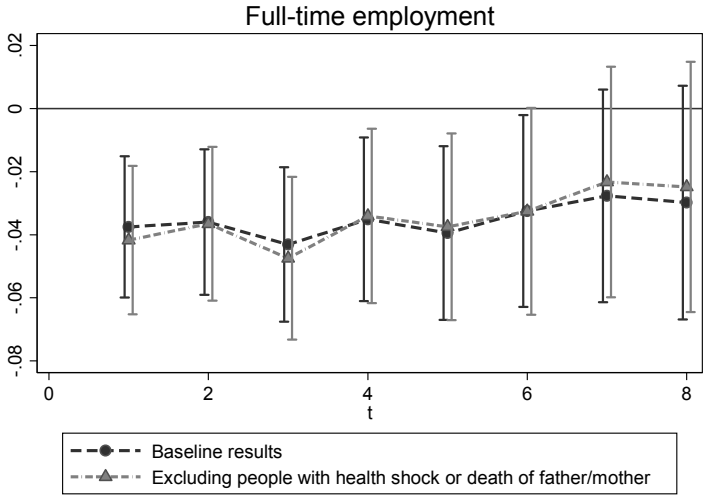


Figure 3.16: Exclusion of individuals with potential health shock or death of a parent between 0 and 1

Source: SOEP. Own calculations. Note: The graph shows the point estimates and the 95% confidence intervals.

Labor market history and expectations

While we control for pre-treatment labor force status and wages which capture a great deal of permanent unobserved heterogeneity and opportunity cost of care, this probably does not perfectly reflect labor market attachment. For instance, last year's employment status does not say a lot about the stability of this job and a low last year's wage might either be low due to a generally low productivity or because of a missed promotion (maybe voluntarily because a careprovision spell was anticipated). Moreover, we do not control for labor market expectations which might affect the willingness to provide care.

In the following specification we also match on labor market participation and hourly wages in the period of five years before the start of the care provision spell (that is, full-time work five years before, full-time work four years before, and so on). This should account for a great deal of labor market dynamics that reflect potential opportunity costs and affect the willingness to provide care. We do not include these variables in our main specification as this longer panel information is not available for each individual in the sample and we would like to maximize sample size.

Moreover, we take a proxy for expectations about the current employment situation into account. This is the answer to the question "Are you concerned about your job stability?" with the possibility to answer "very concerned", "somewhat concerned", or "not concerned at all". Women who expect to lose their job in the near future, might be more likely to provide care. This is not a perfect measure, as it is only available for those who are currently employed and, thus, not included in the baseline specification. Figure 3.17 reports the results when these variables are taken into account. While the coefficients are slightly attenuated, this does not affect the main conclusions at all.

Assuming adverse measurement error

A further scenario in which we would falsely attribute the observed correlation between labor supply and informal care to the effect can arise under presence of non-classical measurement error. Assume a situation where individuals that suffer from unemployment falsely report a positive amount of hours spent caring in order to justify their unemployment. This would inflate our estimates.

By assuming a worst case scenario, we reassign all individuals who stop working between $t = 0$ and $t = 1$ to the control group of non-carers (independent of their reported care status). As the majority of individuals does not state to provide care anyway, this effectively only changes the treatment status of 84 women who stopped working and report a positive amount of hours cared. Yet, this change will mechanically drive our estimates towards zero as we absorb some of the observed correlation that adds up to our baseline effects. The major question is just how strong. Figure 3.18 shows this impact on the results (gray triangles). With such a drastic measurement error where each individuals that gave up her job falsely reported to also provide care, the effects of care provision on full-time work would change to the region of the previously seen lower bound from Section 3.5.1 but remained statistically and economically important.

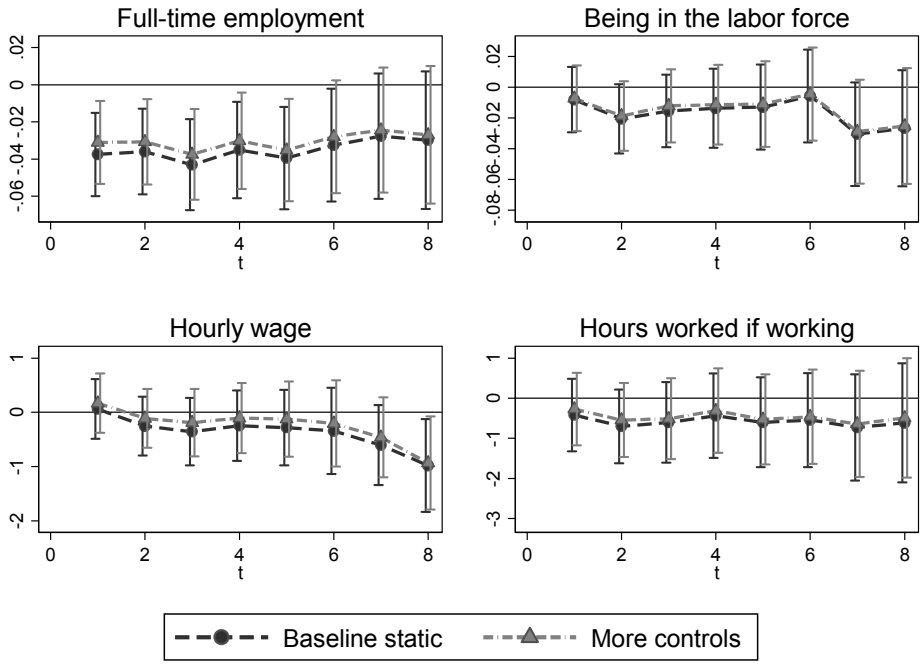


Figure 3.17: Including labor market history and expectations

Source: SOEP. Own calculations. Note: The graph shows the point estimates and the 95% confidence intervals.

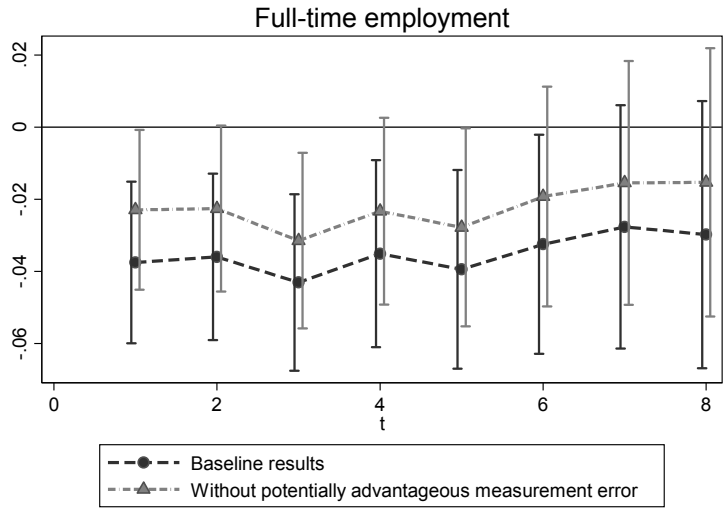


Figure 3.18: Impact of assumed adverse measurement error, full-time work

Source: SOEP. Own calculations. Note: The graph shows the point estimates and the 95% confidence intervals.

Alternative specifications

Figure 3.19 reports results when we split the sample at the median age of 44 in $t = 1$ and carry out analyses for the two groups younger and older than 44. Both for legibility and sample size reasons we restrict the analyses to the static case of providing care in year 1. This can be justified by the small differences between caring for at least one year compared to at least two or three years.

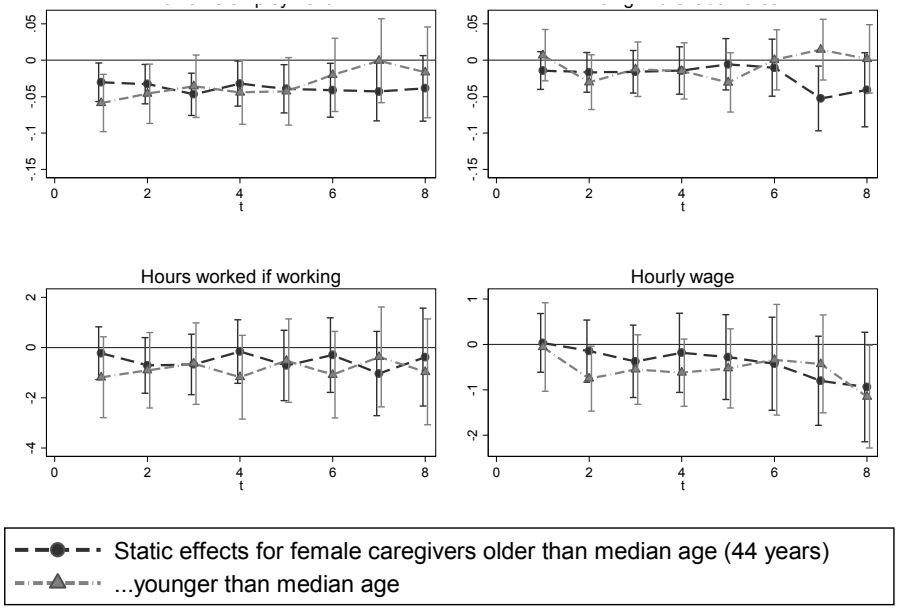


Figure 3.19: Results of the static version – Younger vs. older than median age

Source: SOEP. Own calculations. Note: The graph shows the point estimates and the 95% confidence intervals.

The results are remarkably similar for both groups, with minor differences. The short-run effect on full-time employment is roughly the same for both groups. Yet, after year 5 younger individuals who provided care before are no less likely to work full-time anymore than their no caring counterparts. It is, however, surprising that this drop back to zero appears between year 5 and 6 and we do not have an explanation why this should take place exactly at this point in time. Conditional hours evolve quite similar while the effects on employment are slightly smaller for younger individuals. Finally, short-term wage effects are slightly larger for younger individuals, yet, not significant either.

Selected variables using the double selection procedure

Table 3.6: Variable selection, exemplary case: full-time work, static version

Non-interacted variables

Pre-treatment outcome: full-time
Marital status: Married
Marital status: Widowed
Number of kids in hh
SAS= Very Bad
Conscientiousness
Age of mother $\in [50, 59]$
Age of father $\in [50, 59]$

Interacted variables
((First Variable) $\times$ (Second Variable))

First variable	Second variable
Own age $\in [25, 29]$	Education: Higher
Own age $\in [25, 29]$	Marital status: Widowed
Own age $\in [25, 29]$	SAS= Very Bad
Own age $\in [25, 29]$	Age of father $\in [70, 79]$
Own age $\in [30, 34]$	SAS= Good
Own age $\in [30, 34]$	Doctor visits
Own age $\in [30, 34]$	Conscientiousness
Own age $\in [30, 34]$	Year=2009
Own age $\in [30, 34]$	Age of mother $\in [80, 89]$
Own age $\in [30, 34]$	Age of father $\in [80, 89]$
Own age $\in [35, 39]$	Marital status: Widowed
Own age $\in [35, 39]$	Risk aversion
Own age $\in [35, 39]$	Federal State: HH
Own age $\in [35, 39]$	Federal State: HB
Own age $\in [35, 39]$	Mother info missing
Own age $\in [40, 44]$	Openness
Own age $\in [45, 49]$	Openness
Own age $\in [45, 49]$	Risk aversion
Own age $\in [50, 54]$	Education: Higher
Own age $\in [50, 54]$	Partner existent
Own age $\in [50, 54]$	Conscientiousness
Own age $\in [50, 54]$	Openness
Own age $\in [50, 54]$	Age of father $\in [60, 69]$
Own age $\in [55, 59]$	Mother alive
Education: A-Levels	Federal State: ST
Education: Voc. train.	Age of father $\in [90, 99]$
Education: Higher	Foreign
Education: Higher	Agreeableness
Education: Higher	Age of mother $\in [70, 79]$
Education: Higher	Mother alive
Education: Missing	Federal State: BE
Education: Missing	Federal State: SH
Education: Missing	Federal State: RPSL
Education: Missing	Federal State: ST
Marital status: Married	Partner existent
Marital status: Married	Age of partner $\in [60, 69]$
Marital status: Married	Federal State: NRW
Marital status: Married	Age of mother $\in [70, 79]$
Marital status: Married	Age of mother $\in [80, 89]$
Marital status: Married	Mother alive
Marital status: Divorced	SAS= Good
Marital status: Widowed	Doctor visits
Marital status: Widowed	Reciprocity, positive
Marital status: Widowed	Age of father $\in [50, 59]$
Marital status: Widowed	Age of father $\in [90, 99]$
Foreign	Federal State: SH
Foreign	Federal State: HH
Foreign	Age of father $\in [90, 99]$
Number of kids in hh	Age of partner $\in [50, 59]$
Number of kids in hh	SAS= Good
Number of kids in hh	Hospital stays
Number of kids in hh	Federal State: BW

Interacted variables – continued

First variable	Second variable
Age of partner $\in [60, 69]$	Hospital stays
Age of father $\in [70, 79]$	Mother info missing
Age of father $\in [60, 69]$	Mother alive
Age of father $\in [50, 59]$	Mother info missing
Age of father $\in [50, 59]$	Father alive
Age of mother $\in [80, 89]$	Age of father $\in [60, 69]$
Age of mother $\in [50, 59]$	Mother alive
Federal State: ST	Age of mother $\in [90, 99]$
Federal State: BW	Father info missing
Federal State: RPSL	Age of mother $\in [90, 99]$
Federal State: HE	Age of father $\in [60, 69]$
Federal State: HH	Mother info missing
Federal State: HH	Age of father $\in [60, 69]$
Federal State: HH	Age of mother $\in [60, 69]$
Federal State: HH	Age of mother $\in [50, 59]$
Federal State: SH	Age of mother $\in [90, 99]$
Extraversion	Age of mother $\in [60, 69]$
Openness	Age of father $\in [80, 89]$
Openness	Age of mother $\in [80, 89]$
Openness	Age of mother $\in [60, 69]$
Openness	Age of mother $\in [50, 59]$
Agreeableness	Age of mother $\in [50, 59]$
Conscientiousness	Mother alive
Conscientiousness	Age of mother $\in [70, 79]$
Hospital stays	Age of mother $\in [80, 89]$
Doctor visits	Reciprocity, positive
SAS= Very Bad	Federal State: BB
Age of partner $\in [80, 89]$	Year=2007
Age of partner $\in [70, 79]$	Year=2007
Age of partner $\in [70, 79]$	Year=2004
Age of partner $\in [70, 79]$	Federal State: TH
Age of partner $\in [70, 79]$	Federal State: NI
Age of partner $\in [70, 79]$	Federal State: SH
Age of partner $\in [70, 79]$	Federal State: BE
Partner existent	Age of father $\in [80, 89]$
Partner existent	Doctor visits
Marital status: Single	Father alive
Marital status: Single	Federal State: HH
Marital status: Single	Openness
Marital status: Single	Conscientiousness
Number of kids in hh	Mother alive

Parameters for calibration of the sensitivity analysis

Table 3.7: Parameters for calibration of the sensitivity analysis

Variable	p_{11}	p_{10}	p_{01}	p_{00}	s	d
Pre-treatment outcome:						
full-time	0.89	0.09	0.89	0.07	-0.05	0.82
Age of mother $\in [50, 59]$	0.1	0.08	0.23	0.19	-0.12	0.04
Age of mother $\in [60, 69]$	0.16	0.15	0.23	0.21	-0.07	0.02
Age of mother $\in [70, 79]$	0.27	0.24	0.22	0.18	0.05	0.04
Age of mother $\in [80, 89]$	0.23	0.24	0.08	0.1	0.15	-0.02
Age of mother $\in [90, 99]$	0.02	0.03	0.01	0.02	0.01	-0.01
Mother alive	0.78	0.74	0.78	0.71	0.02	0.07
Mother info missing	0.05	0.04	0.05	0.06	-0.01	-0.01
Age of father $\in [50, 59]$	0.06	0.05	0.16	0.12	-0.09	0.04
Age of father $\in [60, 69]$	0.13	0.11	0.2	0.18	-0.07	0.02
Age of father $\in [70, 79]$	0.19	0.18	0.17	0.15	0.02	0.01
Age of father $\in [80, 89]$	0.16	0.14	0.07	0.07	0.08	0
Age of father $\in [90, 99]$	0.02	0.03	0.01	0.02	0.01	-0.01
Father alive	0.56	0.51	0.62	0.55	-0.05	0.06
Father info missing	0.05	0.05	0.07	0.08	-0.02	-0.01
Partner existent	0.74	0.85	0.67	0.82	0.05	-0.15
Age of partner $\in [50, 59]$	0.61	0.64	0.59	0.65	0	-0.06
Age of partner $\in [60, 69]$	0.12	0.19	0.07	0.15	0.05	-0.08
Age of partner $\in [70, 79]$	0.01	0.01	0.01	0.01	0	-0.01
Age of partner $\in [80, 89]$	0	0	0	0	0	0
Age of partner $\in [90, 99]$	0	0	0	0	0	0
Number of siblings	0.56	0.61	0.59	0.63	-0.02	-0.03
Number of brothers	0.61	0.63	0.59	0.63	0.01	-0.04
Own age $\in [25, 29]$	0.05	0.03	0.13	0.08	-0.06	0.05
Own age $\in [30, 34]$	0.03	0.05	0.12	0.11	-0.07	0.01
Own age $\in [35, 39]$	0.08	0.09	0.12	0.14	-0.05	-0.02
Own age $\in [40, 44]$	0.14	0.14	0.16	0.15	-0.02	0
Own age $\in [45, 49]$	0.23	0.18	0.17	0.13	0.05	0.03
Own age $\in [50, 54]$	0.24	0.19	0.15	0.12	0.07	0.03
Own age $\in [55, 59]$	0.18	0.18	0.1	0.12	0.07	-0.02
Education: A-Levels	0.07	0.06	0.09	0.07	-0.02	0.01
Education: Voc0. train0.	0.08	0.08	0.07	0.07	0.01	0.01
Education Higher	0.35	0.17	0.32	0.16	0	0.16
Education: Missing	0.01	0.02	0.01	0.02	0	0
Marital status: Single	0.16	0.07	0.31	0.12	-0.09	0.18
Marital status: Married	0.61	0.78	0.5	0.73	0.09	-0.23
Marital status: Divorced	0.15	0.07	0.08	0.13	0	0.05
Marital status: Widowed	0.03	0.05	0.02	0.03	0.01	-0.01
Foreign	0.03	0.05	0.06	0.09	-0.03	-0.04
Number of kids in hh	0.17	0.34	0.23	0.47	-0.09	-0.24
Neuroticism	0.47	0.51	0.42	0.47	0.05	-0.05
Conscientiousness	0.52	0.46	0.54	0.46	-0.02	0.08
Agreeableness	0.53	0.49	0.48	0.46	0.03	0.02
Openness	0.59	0.55	0.59	0.53	0.01	0.06
Extraversion	0.63	0.61	0.61	0.6	0.01	0.01
Reciprocity, positive	0.55	0.55	0.54	0.52	0.02	0.02
Reciprocity, negative	0.52	0.52	0.52	0.53	-0.01	-0.01
Risk aversion	0.61	0.52	0.57	0.51	0.02	0.06

Continued on next page

Table 3.7 – continued

Variable	p_{11}	p_{10}	p_{01}	p_{00}	s	d
SAS= Very Good	0.09	0.06	0.11	0.08	-0.02	0.03
SAS= Good	0.36	0.38	0.49	0.43	-0.08	0.05
SAS= Satisfactory	0.41	0.36	0.3	0.32	0.06	-0.02
SAS= Bad	0.13	0.16	0.09	0.13	0.04	-0.04
SAS= Very Bad	0.01	0.03	0.01	0.03	0	-0.02
Doctor visits	0.32	0.39	0.3	0.37	0.03	-0.07
Hospital stays	0.08	0.11	0.08	0.14	-0.01	-0.05
Federal State: BE	0.07	0.03	0.05	0.03	0	0.02
Federal State: SH	0.02	0.03	0.02	0.03	0	-0.01
Federal State: HH	0.01	0.01	0.02	0.01	0	0
Federal State: NI	0.06	0.09	0.08	0.09	0	-0.02
Federal State: HB	0	0.01	0.01	0.01	0	0
Federal State: NRW	0.18	0.21	0.18	0.22	0	-0.04
Federal State: HE	0.06	0.09	0.07	0.07	0.01	0
Federal State: RPSL	0.07	0.06	0.05	0.07	0	-0.02
Federal State: BW	0.07	0.13	0.11	0.13	-0.01	-0.02
Federal State: BY	0.1	0.14	0.14	0.15	-0.02	-0.01
Federal State: BB	0.07	0.04	0.05	0.03	0.01	0.01
Federal State: ST	0.07	0.03	0.06	0.03	0	0.02
Federal State: TH	0.05	0.04	0.05	0.04	0	0.01
Federal State: SN	0.13	0.07	0.09	0.06	0.02	0.03

Notes: Source SOEP

Additional material

Table 3.8: Back-of-the-envelope calculation of fiscal effects

Mean employment of never-carers		Estimated effects due to informal care (averaged over all eight periods)		
Hourly wage	14.23	Wage effect:	-0.38	13.85
Cond. hours	32.24	Eff. on hours:	-0.58	31.66
Employed:	0.77	Employment effect:	-0.017	0.75

Fiscal effects:				
	Prior to informal care	Due to informal care		
Uncond. hours:	24.82	23.84		
Wage differential:			Difference:	
Labor income per year	18,228.06	17,037.48	-1,190.58	
Tax differential:				
Average income tax rate:	12.23%	11.28%		
thereof income tax:	2,229	1,922	-307	
Average consumption rate:	0.931			
Value-added-tax (VAT) rate (weighted average between 7 and 19%)	0.15			
Total amount VAT:	2,234.23	2,110.90	-123.33	
		Total	-430.33	
		# female informal caregiver:	2m	
			-860.66m €	less tax revenue p.a.

In order to calculate back-of-the-envelope fiscal effect, we start with the counterfactual average level of labor supply (Employment probability and conditional hours) and the respective mean wage of non-carer and add the estimated effects to get levels in labor supply and the wage for caregivers. Now we can calculate gross incomes for caregivers and non-carers. Now we can roughly estimate tax differential between both groups. In order to compute the income tax, we make use of the average tax rate for both incomes. Additionally, the VAT also contributes to the tax differential and we calculate this based on an average consumption rate. All in all caregivers annually pay €473 less taxes. Multiplied with 2 million female caregivers in Germany, this would translate into an annual fiscal loss of €860.33m due to informal caregiving.

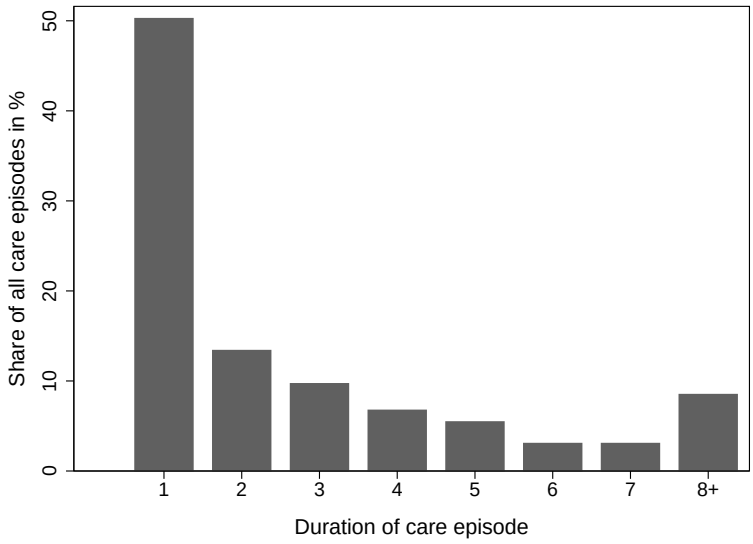


Figure 3.20: Distribution of imputed care episodes

Source: SOEP, own calculations. This graph shows the distribution of episodes of consecutive care of at least one hour per day. The data are restricted to starting waves 1 to 5 as defined in Figure 3.1 to ensure that every spell can last for at least 8 years. Imputation is according to the following scheme: whenever a women provided care in a certain year and two years later, the care indicator one year later is set to 1 even if no care provision is stated in the questionnaire.

Common support restrictions

Figure 3.21 shows the distribution of the propensity scores on each node of the decision tree (see Figure 3.8) by actual care state (light gray for caregivers, dark gray for non-carers) – exemplarily for full-time work as an outcome and the first two periods. The red vertical lines indicate the region of common support (the smallest set of overlap in the support between both groups). The upper panel depicts this observed probability of starting caregiving after the initial period ($t = 0$). The projected probabilities are small, reflecting the low fraction of caregivers in period $t = 1$. The middle panel shows the same for the second period (caregiving in $t = 2$) conditional on having cared in $t = 1$, here the odds are balanced. The bottom panel plots the propensity score of continuing not to care for the second period $t = 0$. All those propensity scores are used to construct the inverse probability weights for the dynamic estimates. The overall conclusion from this graph is, that the overlap between treatment and control group is good and that the restriction on the common is not crucial. Out of the 63,372 observations we drop 710 at the first node, 138 at the second, and 641 at the third which are off the common support. The restriction on observations lying within the 5^{th} and 95^{th} quantile is more binding where we drop 6,337 observations.

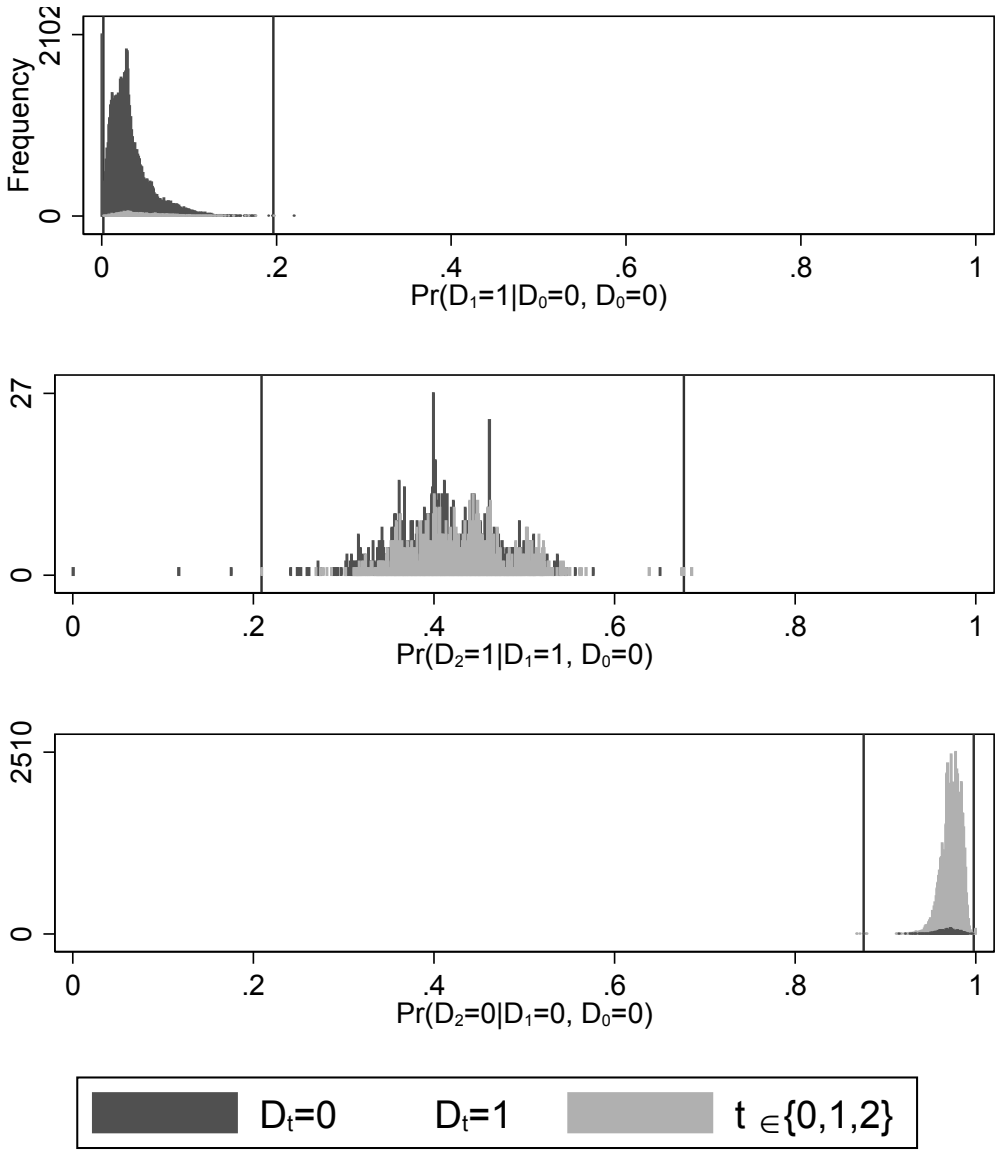


Figure 3.21: Common support

Part III

Consequences of the educational expansion

Chapter 4

Heterogeneity in marginal non-monetary returns to higher education[1]

4.1 Introduction

"The whole world is going to university – Is it worth it?" The Economist's headline read in March 2015.[2] 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., 2006a, 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

[1] This chapter is written jointly with Daniel Kamhöfer and Hendrik Schmitz and is published as: Kamhöfer, D. A., Schmitz, H., and Westphal, M. (2017). Heterogeneity in Marginal Non-monetary Returns to Higher Education. Journal of the European Economic Association, forthcoming. Funding by the German Research Foundation (DFG, Grant number SCHM 3140/1-1) is gratefully acknowledged.

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

the compound manifestation of the overall effect has rarely been studied for college education over a short-term horizon[3] 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[4] 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 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

[3] 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.

[4] See Section 4.4 for a detailed definition of cognitive abilities. We use the terms "cognitive abilities", "cognitive skills", and "skills" interchangeably.

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% of individuals with the lowest desire to study. Mental health effects are zero throughout the population. Thus, our 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 4.2 briefly introduces the German educational system and describes the exogenous variation we exploit. Section 4.3 outlines the empirical approach. Section 4.4 presents the data. The main results are reported in Section 4.5 while Section 4.6 addresses some of its potential underlying pathways. Section 4.7 concludes.

4.2 Institutional background and exogenous variation

4.2.1 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[5] and universities of applied science (Fachhochschulen). The regular time to receive

[5]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).

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.

4.2.2 Exogenous variation in college education over time

While the higher educational system as described in Section 4.2.1 did not change in the years under review, the accessibility (in terms of mere quantity but also distribution 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).[6] 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.[7] 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 4.1. Figure 4.6 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

[6]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.

[7]All data are taken from the German Statistical Yearbooks, 1959-1991, see German Federal Statistical Office (various issues, 1959–1991). We only use colleges and no other higher educational institutes described in Section 4.2 (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.

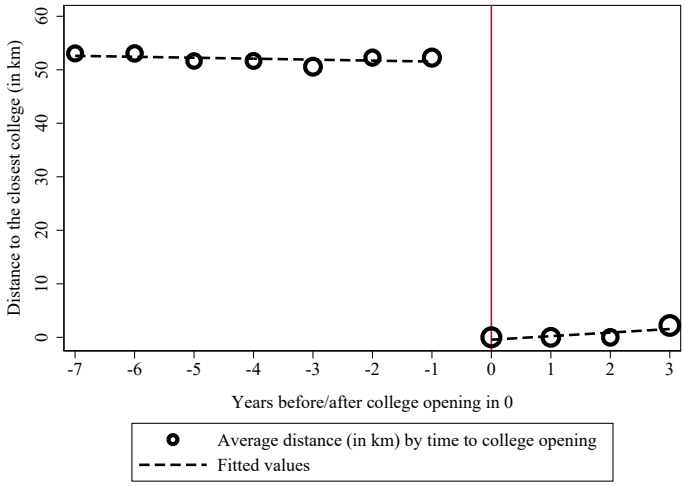


Figure 4.1: Average distance to the closest college over time for districts with a college opening

Notes: Own illustration. Information on colleges are taken from the German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, various issues, 1959–1991). 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.

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 4.2 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 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

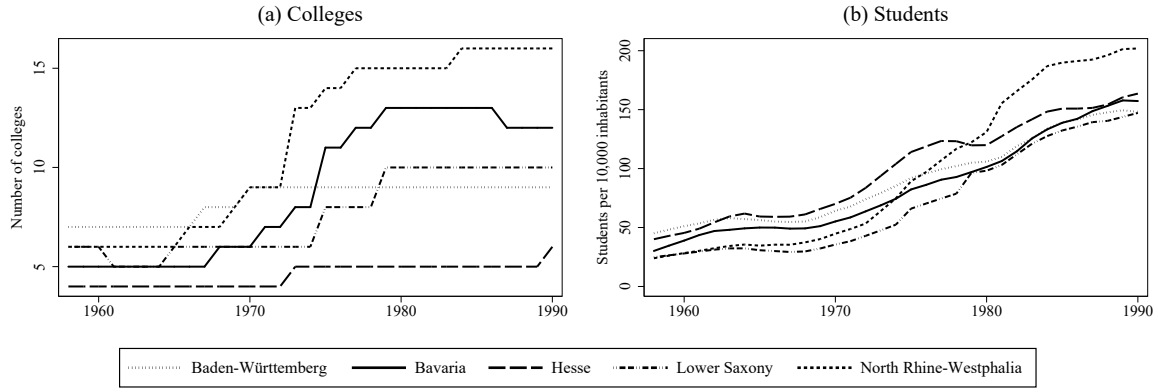


Figure 4.2: Number of colleges and students over the time in selected states

Notes: Own illustration. College opening and size information are taken from the German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, various issues, 1959–1991). 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.

openings within the federal state. See the Supplementary Materials A to the paper for a much more detailed description of the political process involved.

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 4.1 reports some numbers on the regional level as of the year 1962 (the earliest possible year available to us with representative data).[8] 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 socio-economic controls (including the socioeconomic environment before the college decision, see Section 4.4), 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 4.1 shows a decline in the population density by 6% 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% 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 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.

[8] Table 4.1 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 4.1: 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.

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 4.4.4 after presenting the data.

4.3 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:

$$Y^1 = X'\beta_1 + U_1 \tag{4.1}$$
$$Y^0 = X'\beta_0 + U_0 \tag{4.2}$$
$$D^* = Z'\delta - V \qquad \text{where } D = \mathbf{1}[D^* \geq 0] = \mathbf{1}[Z'\delta \geq V] \tag{4.3}$$

The vector X contains observable, and U_1, U_0 unobservable factors that affect the potential outcomes.[9] 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 $Z'\delta$

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

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).[10]

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 (2002) showed its equivalence.[11] 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. (2006b) 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

[10] 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 $P(Z) \equiv P(D = 1|Z) = \Phi(Z'\delta)$.

[11] 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.

as it unfolds all local switching effects by intrinsic 'willingness' to study and not only some weighted average of those.[12]

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 in (4.1) and (4.2) 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:

$$\begin{aligned} E(Y|X,p) &= X'\beta_0 + X'(\beta_1 - \beta_0)\cdot p + E(U_1 - U_0|D=1,X)\cdot p \\ &= X'\beta_0 + X'(\beta_1 - \beta_0)\cdot p + K(p) \end{aligned} \tag{4.4}$$

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 (4.4) 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 (4.1) and (4.2):

$$\frac{\partial E(Y|X,p)}{\partial p} = X'(\beta_1 - \beta_0) + \frac{\partial K(p)}{\partial p} \tag{4.5}$$

To achieve non-parametric identification of the terms in Equation (4.5), 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 (4.5) unconditional independence is required: $(U_1, U_0, V) \perp\!\!\!\perp Z, X$ (Carneiro et al., 2011, for more details consult the Supplementary Materials).[13]

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

[12] 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.

[13] 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.

estimating the overall level and the curvature of the MTE. Our solution 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 (4.4) by restricting $\beta_1 = \beta_0 = \beta$ except for the intercepts α_1, α_0 in (4.1) and (4.2) such that $E(Y|X,p)$ becomes:

$$E(Y|X,p) = X'\beta + (\alpha_1 - \alpha_0)\cdot p + K(p) \tag{4.6}$$

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 (4.4) 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 (4.6) 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.[14]

In estimating (4.6), we follow Carneiro et al. (2010, 2011) again and use semi-parametric techniques as suggested by Robinson (1988).[15] Standard errors are clustered at the district level and were generated by bootstrapping the entire procedure using 200 replications.

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

[15] 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 $K(p)$ 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 $\widehat{MTE}(x,p) = (\widehat{\alpha}_1 - \widehat{\alpha}_0) + \sum_{j=1}^{k}\widehat{\phi}_j j p^{j-1}$.

4.4 Data

4.4.1 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. (2011a). 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 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.

4.4.2 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 4.2 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 4.7 in the Appendix. In the regression analysis we use log gross hourly wages.

Table 4.2: 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).

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).[16] 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 deviation 1, where higher values indicate better health. Columns (2) to (3) of Table 4.2 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., 2007, for a comparison of cognitive skill tests in the German Socioeconomic 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.[17] 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.[18] 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

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

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

[18] 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).

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 4.2. 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).[19]

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.

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).

4.4.3 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 4.3). 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).

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

Table 4.8 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 4.2), 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).

4.4.4 Instrument

The processes of college expansion discussed in Section 4.2.2 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:

$$Z_{it} = \sum_{j}^{326} K(dist_{ij}) \times \left(\frac{\#students_{jt}}{\#inhabitants_{jt}} \right). \tag{4.7}$$

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 4.3 presents the descriptive statistics. We also provide background information on certain descriptive measures on distance and student density.

Table 4.3: 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, various issues, 1959–1991). Distances are calculated as the Euclidean distance between two respective district centroids.

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. 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

$$Z_{it} = \sum_{j}^{326} K(dist_{ij}) \times \mathbb{1}[\text{college avaiable}_{jt}] \tag{4.8}$$

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 4.8 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.

4.5 Results

4.5.1 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 4.4, 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 % 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 4.2, slightly attenuated, however, due to the inclusion of control variables.

Table 4.4: 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 4.4 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 (4.7), 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.

4.5.2 Marginal treatment effects

Figure 4.3a 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 4.3b. 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 semi-parametric 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.

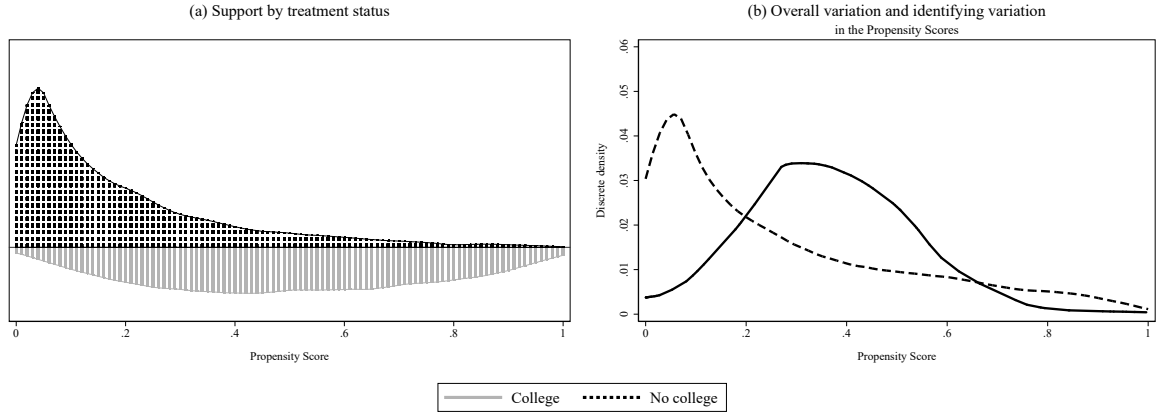


Figure 4.3: Distribution of propensity scores

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.[20] 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 (4.6) in Section 4.3).

[20] 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 4.4 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 4.3. The returns are close to 80% 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% of individuals with the highest desire to study, while there is no clear evidence for any returns for the remaining 30%. 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% per year of college, while we find 80% 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 4.4 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 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 4.4. 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% of the individuals while they are close zero for the 25% 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% of individuals. This can also be read as suggestive evidence for cognitive abilities being a channel for the effect of education on wages.

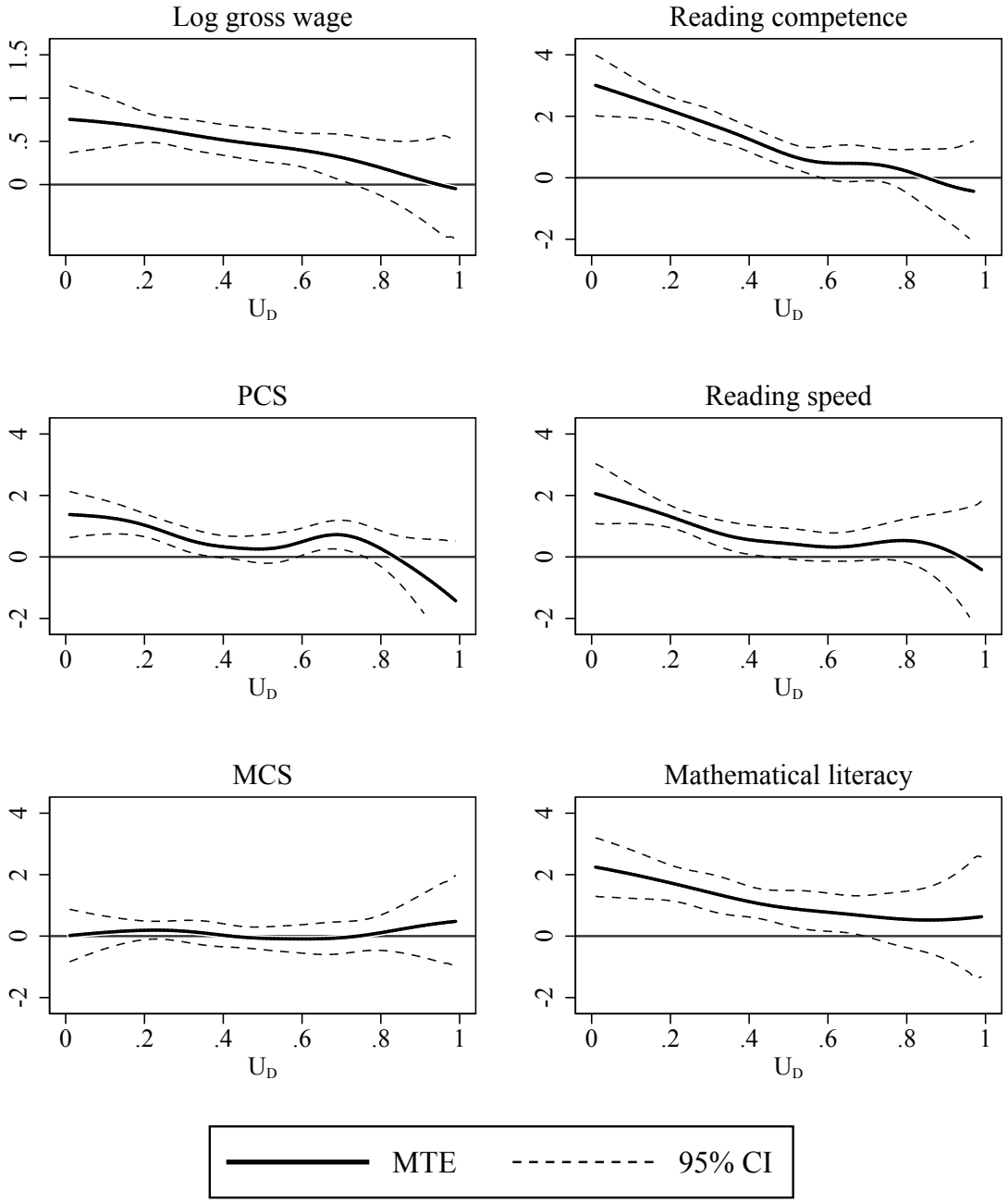


Figure 4.4: Marginal Treatment Effects for cognitive abilities and health

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% 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.

- 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 4.4.4). Figure 4.8 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 (4.7) seems somewhat arbitrary, these robustness checks show that the specification of the instrument does not affect our conclusions.

4.5.3 Treatment parameters

Table 4.5 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. (2006b). 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 4.5. 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 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.[21] As depicted in Figure 4.3, 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% without college degree (and arbitrarily close to 0% with a degree). Instead, the parameters in Table 4.5 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.

Table 4.5 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 4.5 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

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

Table 4.5: 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 semi-parametric Robinson estimator. The LATE is estimated using the IV weights depicted in Figure 4.5. 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.

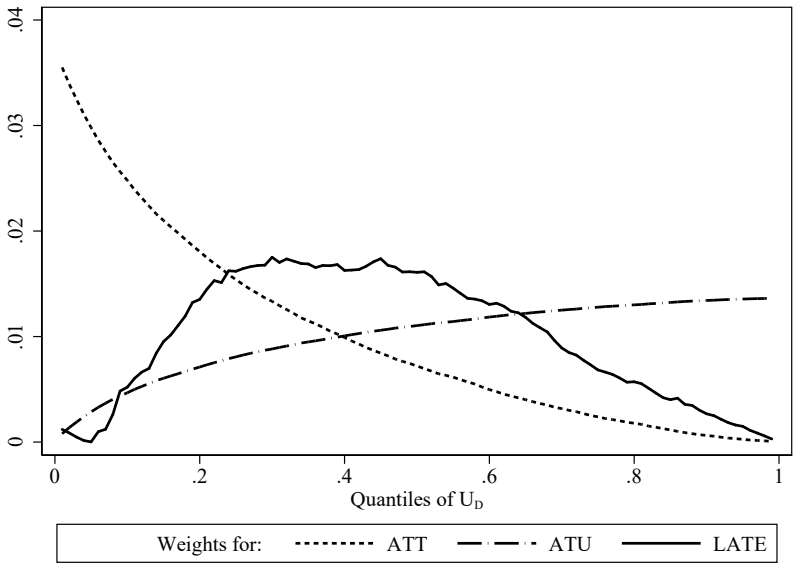


Figure 4.5: Treatment parameter weights conditional on the propensity score

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.

mathematical literacy and reading competence are about 25% of a standard deviation for all parameters. For reading speed the effects are around 15% of an SD, while

the wage effects are around 10%. 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 ($\Phi(0) = 50\%$), the average effect of 1.11 of a standard deviation, all other things the same, will make you end up at the 87% quantile of that distribution ($\Phi(0 + 1.11) = 87\%$) – 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 4.4, the health returns to higher education are smaller than the skill returns, still they are around 10% 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% level would increase from 1.96 to 2.65 and would not change any conclusions regarding significance.[22]

4.6 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;
- 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).

[22] 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.

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 4.6. 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.

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.

Table 4.6: Potential mechanisms for cognitive skills

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

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 4.7 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 4.7 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

Table 4.7: Potential mechanisms for health

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

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.

4.7 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% 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.

4.8 Appendix

Figures

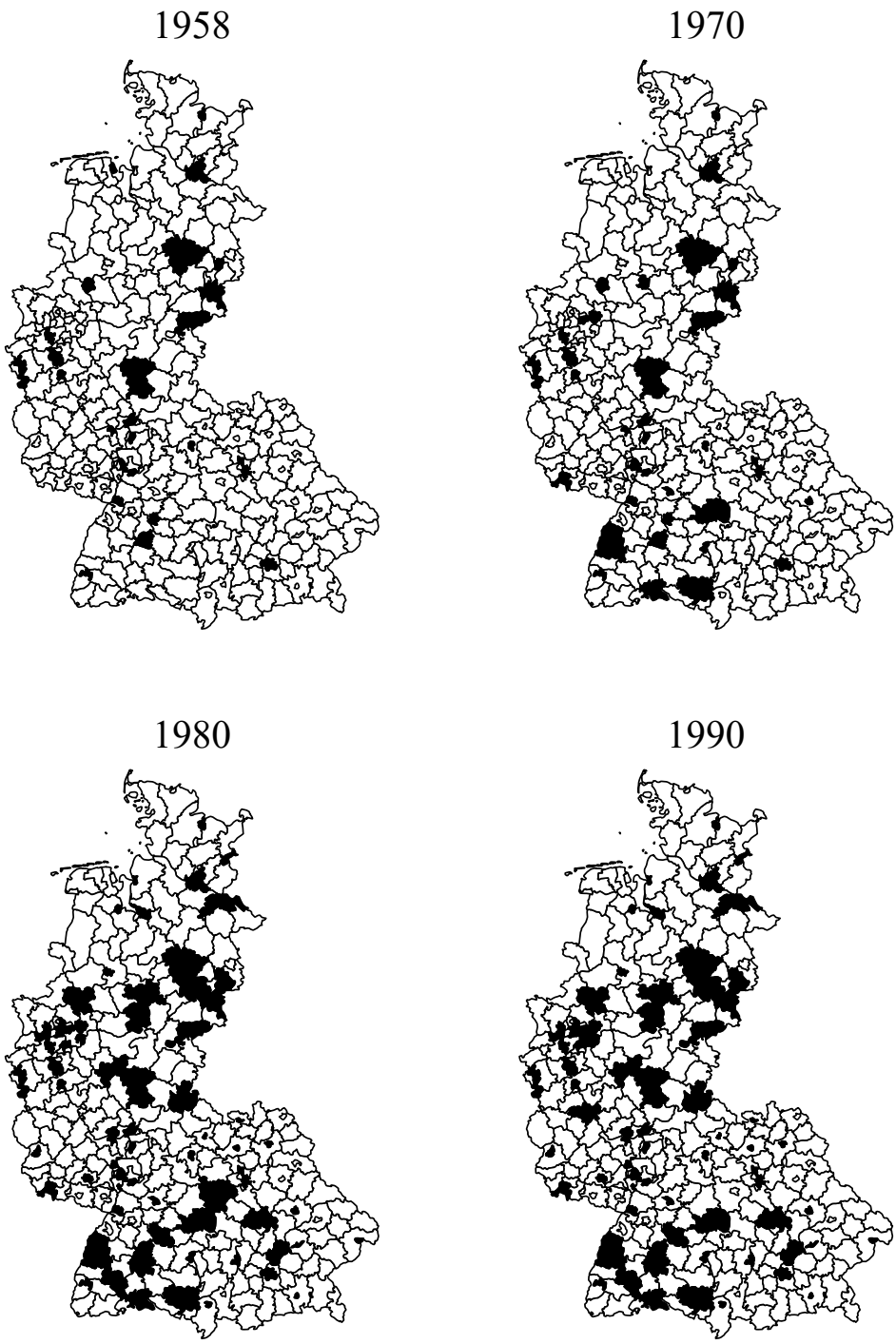


Figure 4.6: Spatial variation of colleges across districts and over time

Notes: Own illustration based on the German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, various issues, 1959–1991). 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.

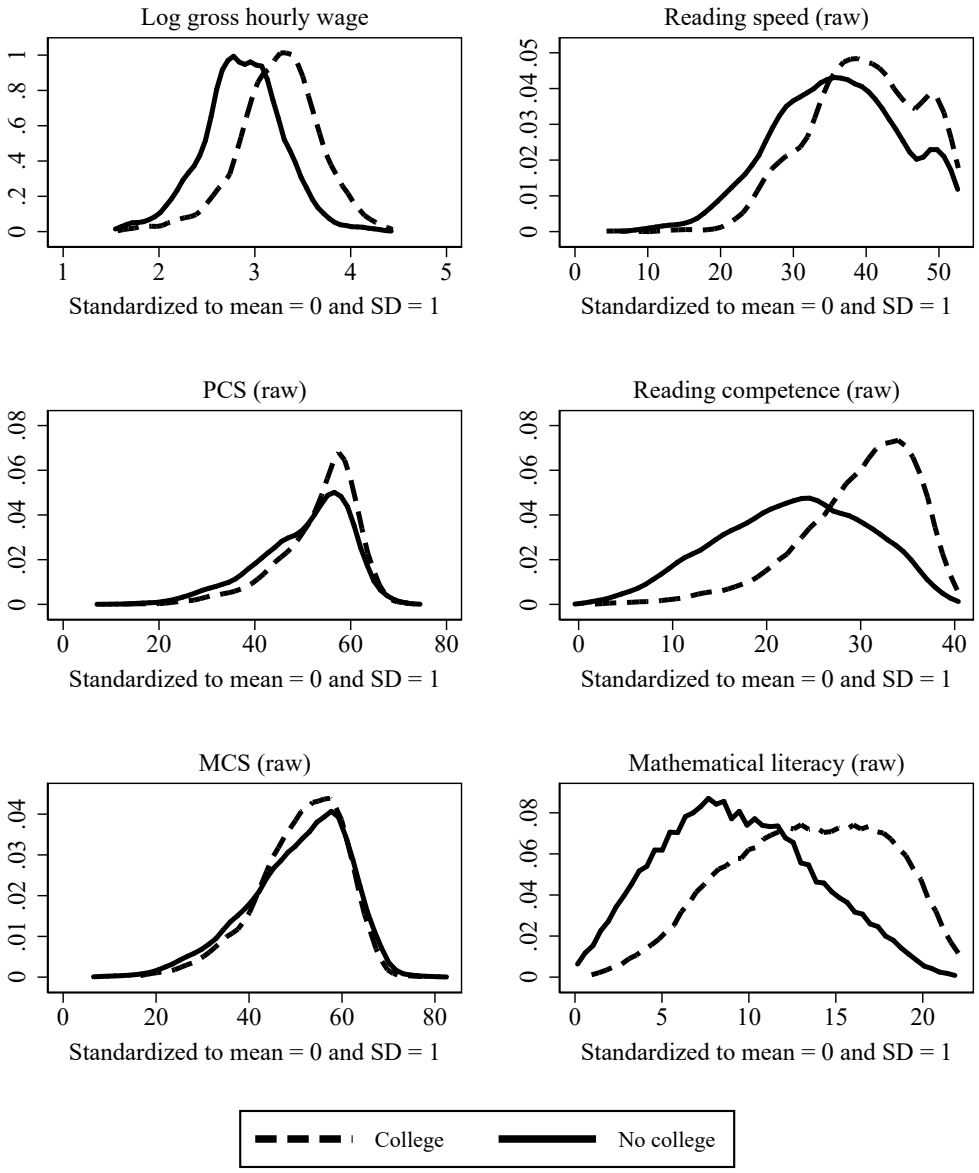


Figure 4.7: Distribution of dependent variables by college graduation

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

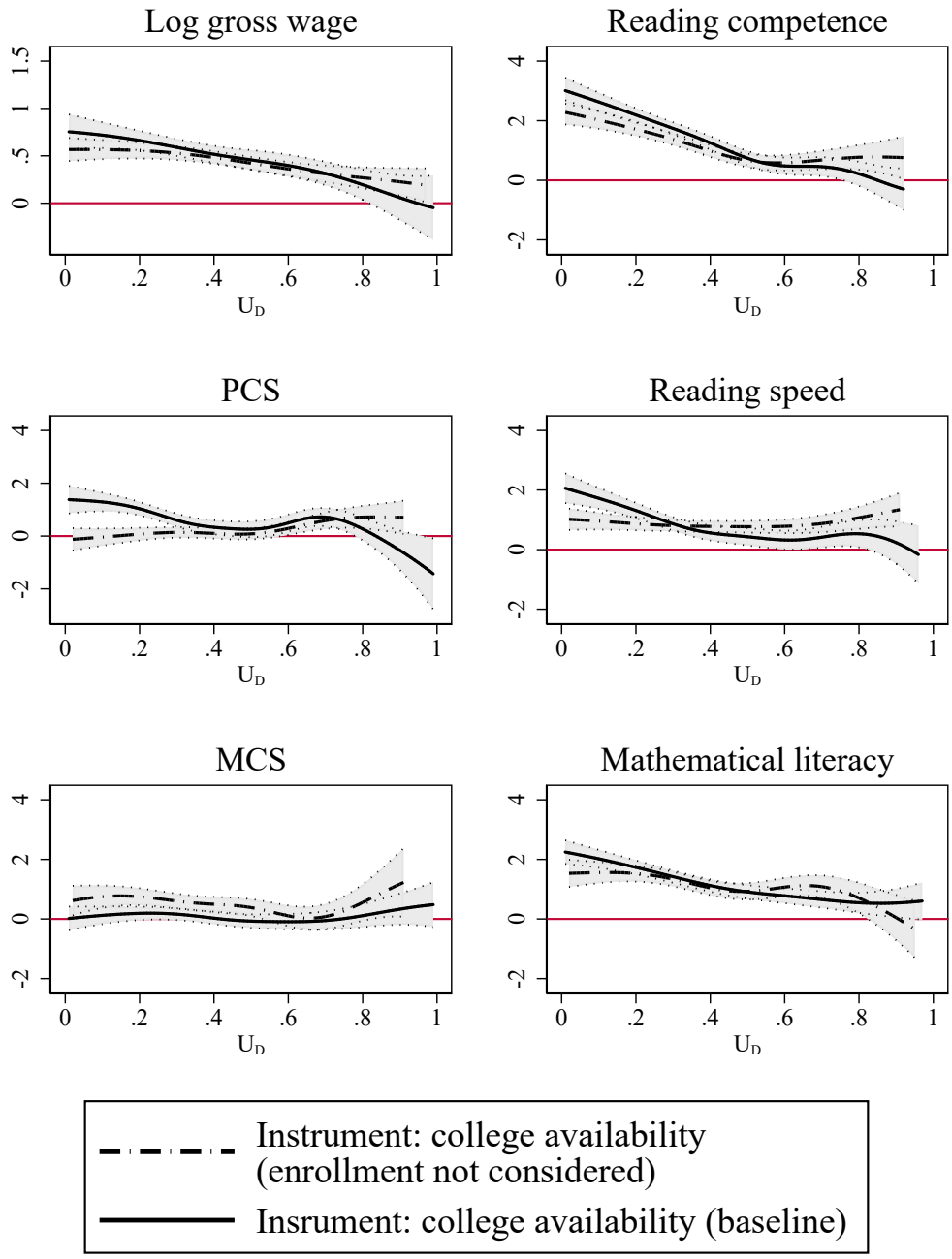


Figure 4.8: Sensitivity in Marginal Treatment Effects when using only the sum of the kernel weighted college distances

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 4.8: 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
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

Continued on next page

Table 4.8 – continued

Variable	Definition	Respondents	
		with college degree	w/o college degree
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.

Supplementary materials

Additional information on the instrument

More information on the process of college openings

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.[23] While the Council of Science and Humanities suggested to link college openings to the expected increase in the population (NRW, 1971b), 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, 1971c, 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, 1971a, 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, 1971a). 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).

[23] 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 4.9: 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)
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

Continued on next page

Table 4.9 – continued

	(1)	(2)	(3)	(4)	(5)	(6)
	(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)
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 4.10: 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)

Continued on next page

Table 4.10 – continued

	(1)	(2)	(3)	(4)	(5)	(6)
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
	(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)
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***

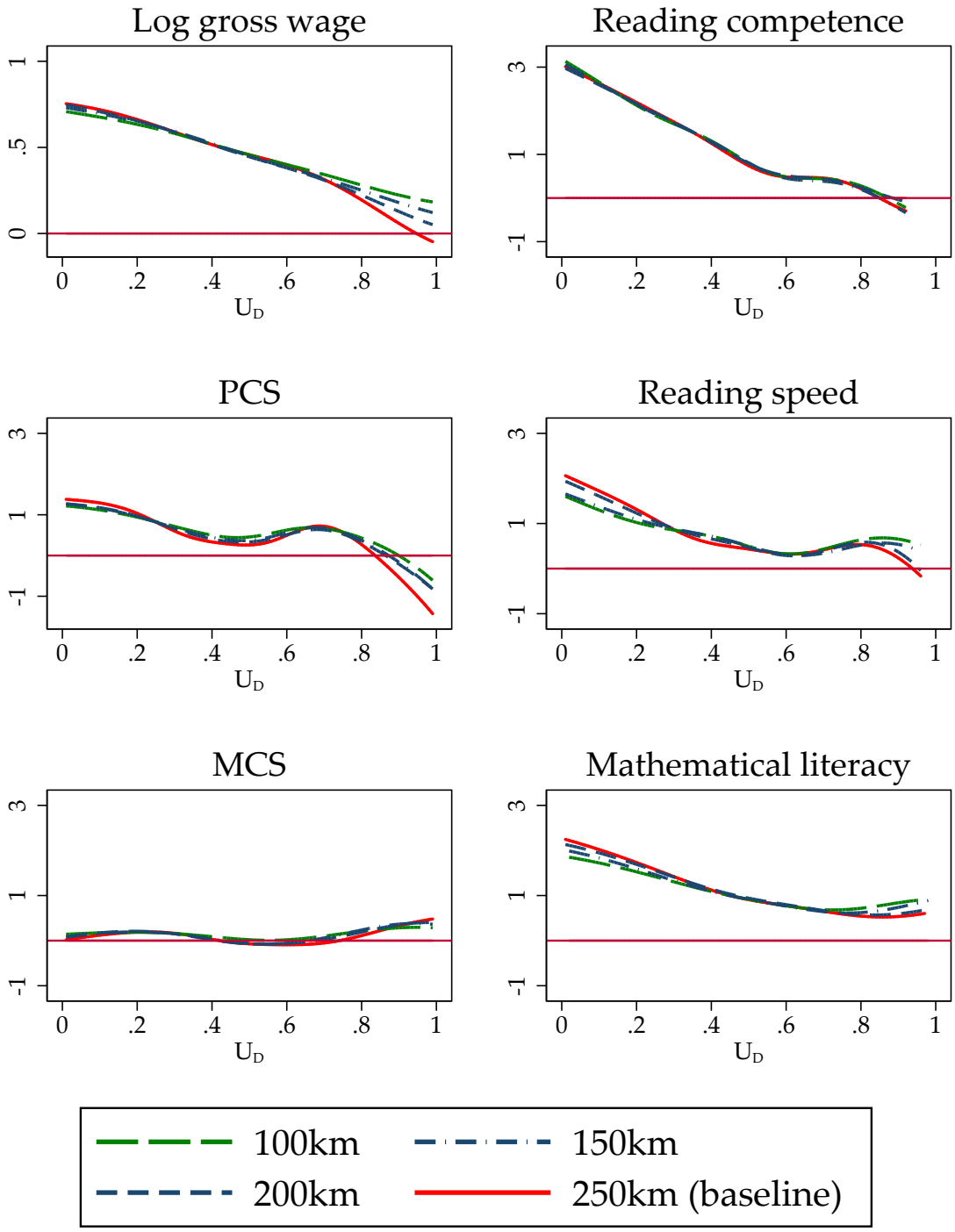
Continued on next page

Table 4.10 – continued

	(1)	(2)	(3)	(4)	(5)	(6)
	(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 4.9: 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 4.11: 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 $E(.|X,Z)$:

$$\begin{aligned} E(Y|X,Z) &= \underbrace{E\left[X\beta_0 + U_0|X,Z\right]}_{\text{CIA: Independent of } Z} + \underbrace{E\left[\left(X(\beta_1-\beta_0) + (U_1-U_0)\right)D|X,Z\right]}_{\text{Law of Iterated Expectations}} \\ &= X\beta_0 + \underbrace{E(U_0|X)}_{=0} + \overbrace{E\left[X(\beta_1-\beta_0) + (U_1-U_0)\,|D=1,X,Z\right]}^{\text{CIA: independent of } Z}\underbrace{E(D|X,Z)}_{=p} \\ &= X\beta_0 + E\left[X(\beta_1-\beta_0) + (U_1-U_0)\,|D=1,X,Z\right]p \\ &= X\beta_0 + X(\beta_1-\beta_0)p + \underbrace{\underbrace{E\left[(U_1-U_0)\,|D=1,X\right]}_{\neq 0}p}_{\substack{\text{Cannot be separated in estimation: one term would} \\ \text{need to be restricted by some assumption.}}} \end{aligned}$$

Under the CIA $X(\beta_1-\beta_0)$ and $E\left[(U_1-U_0)\,|D=1,X\right]$ 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 $E\left[(U_1-U_0)\,|D=1\right]$ needs to be identified and $E\left(X(\beta_1-\beta_0)\right)$ 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 5

Fertility effects of college education: evidence from the German educational expansion[1]

5.1 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.[2] By analyzing the 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.

[1]This chapter is jointly written with Daniel Kamhöfer and published as: Kamhöfer, D. A., and Westphal, M. (2017). Fertility Effects of College Education: Evidence from the German Educational Expansion. Ruhr Economic Papers 717, RWI Essen. Financial support from the German Research Foundation (DFG, Grant number SCHM 3140/1-1) is gratefully acknowledged.

[2]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 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.

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.[3] 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, 2017, 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[4] 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 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

[3] 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.

[4] 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.

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, 2017) 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, 2017).

The remainder of the paper is as follows: Section 5.2 briefly presents the general trends in fertility and higher education in Germany. Section 5.3 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 5.4. The main results on quantitative fertility effects are presented in Section 5.5. Subsequently, Section 5.6 sheds light on the timing and socioeconomic factors that potentially shape the detected fertility patterns before Section 5.7 concludes.

5.2 Trends in fertility and education in Germany

Using official statistics for the whole population, Figure 5.1 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 5.1 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.

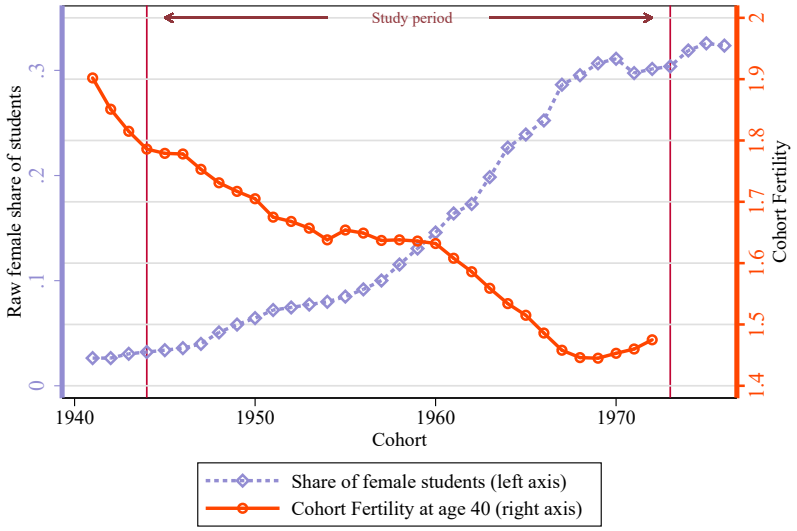


Figure 5.1: Trends in fertility and college enrollment by birth cohort in Germany

Notes: Own calculations using data from Max Planck Institute for Demographic Research and Vienna Institute of Demography (2014) and German Federal Statistical Office (2016). 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.

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 age at the time of the first marriage as depicted in Figure 5.2. 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 5.1, Figure 5.2 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 5.2, 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 5.2 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

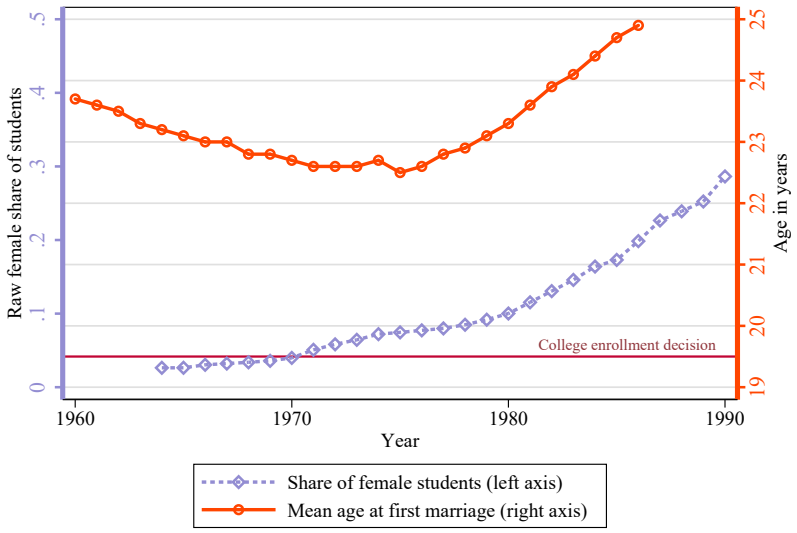


Figure 5.2: Mean age at first marriage and college enrollment by year in Germany

Notes: Own calculations using data from Max Planck Institute for Demographic Research and Vienna Institute of Demography (2014); German Federal Statistical Office (2016). 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.

5.2 suggests that the share of women who took care of their own subsistence (through working for pay or student loans introduced in 1971) increased over time.

5.3 The college expansion

5.3.1 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.[5] 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 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.

[5]The general description of education in Germany and the college expansion is closely related to Kamhöfer et al. (2017) and has been adjusted for the purpose of the analysis conducted here.

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 5.3a). The number of female students in total in the colleges in the sample in Figure 5.3b is similar to the corresponding number in Figure 5.1. 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 5.6 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 5.3b, 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.

5.3.2 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

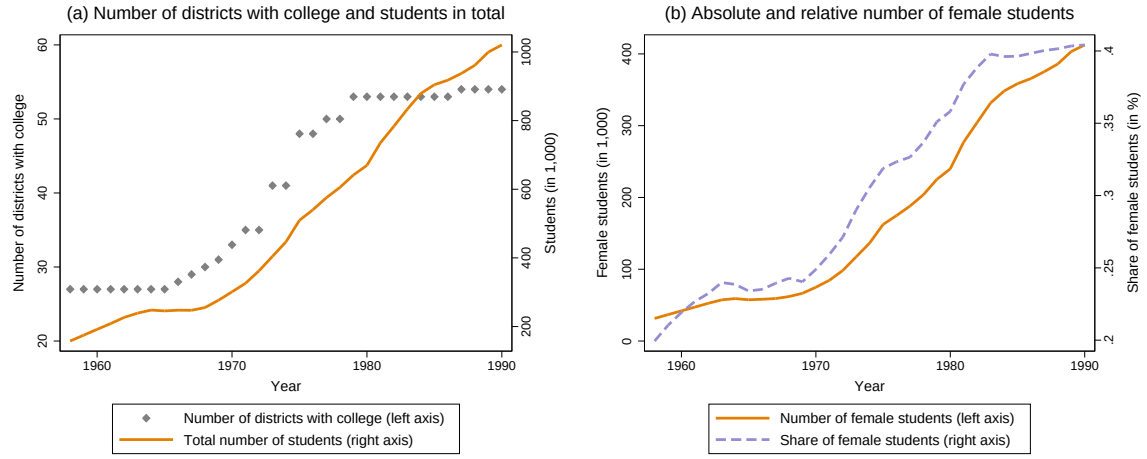


Figure 5.3: Colleges and students over time and by gender

Notes: Own illustration. College opening and size information are taken from the German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, various issues, 1959–1991). 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.

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).[6] 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 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.[7]

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

[6] 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.

[7] Figure 5.7 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.

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 5.8 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 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 (1971b,c). 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 (1971a).

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 5.1 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 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 5.1, 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.

5.4 Data and empirical strategy

5.4.1 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. (2011a).[8] NEPS data map the educational trajectories of more than 60,000 individuals in total. The data set consists of a

[8]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.

Table 5.1: 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		
Potential college determinant	Mean	Mean	Diff.	OLS
Number of kids per capita (total population)	10.497 (0.522)	10.437 (0.283)	−0.15 (0.121)	−0.033 (0.052)
...students	0.016 (0.019)	0.011 (0.011)	−0.008* (0.004)	−10.723 (10.653)
...divorced	0.023 (0.069)	0.017 (0.006)	−0.005 (0.016)	−1.00 (40.185)
...widowed	0.088 (0.015)	0.091 (0.008)	0.007** (0.003)	20.035 (20.357)
...females	0.525 (0.041)	0.528 (0.013)	0.002 (0.01)	−20.918 (10.851)
...migrational background	0.021 (0.022)	0.018 (0.017)	−0.006 (0.005)	−10.698 (10.545)
...unemployed	0.002 (0.001)	0.002 (0.001)	0.001** (0.00)	250.484 (190.743)
Sectoral composition of employment				
- primary	0.029 (0.055)	0.046 (0.053)	0.023* (0.013)	0.39 (0.497)
- secondary	0.543 (0.088)	0.551 (0.069)	0.008 (0.02)	0.147 (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, various issues, 1959–1991). 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; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

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)[9], 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 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 5.7 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 5.2 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 5.2 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 5.3 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 instruments.[10] This is especially the case as we observe college openings. Using, for instance, a scalar for the

[9] For every individual we use only the most recent observation.

[10] 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.

Table 5.2: 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$\geq$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.

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:

$$Z_{it} = \sum_{j}^{325} K(dist_{ij}) \times \left(\frac{\#students_{jt}}{\#inhabitants_{jt}} \right). \tag{5.1}$$

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 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 5.8 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).[11]

5.4.2 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 $\boldsymbol{X}'_{itd}$:

$$Y_{itd} = \beta_0 + \beta_1 D_{itd} + \boldsymbol{X}'_{itd}\beta_2 + u_{itd}. \tag{5.2}$$

In order to separate the general trend in college education from the reverse trend in fertility (as depicted in Figure 5.1), 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 5.7 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 that are already established before college, OLS would overestimate the true college effect.[12] On the other hand, OLS may underestimate the true effect if factors such as the family's wealth are omitted from the model.[13] Also, general preferences for having a family do not necessarily lead to an overestimation of OLS, as females with these preferences may

[11] For alternative specifications of the instrument, see Kamhöfer et al. (2017).

[12] 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.

[13] 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.

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. 5.1 as an instrumental variable in a two-stage least-squares (2SLS) approach. The first stage of the 2SLS approach reads:

$$D_{itd} = \delta_0 + \delta_1 Z_{td} + \boldsymbol{X}'_{itd}\boldsymbol{\delta}_2 + v_{itd}. \tag{5.3}$$

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}:

$$Y_{itd} = \beta_0 + \beta_1 \widehat{D}_{itd} + \boldsymbol{X}'_{itd}\beta_2 + u_{itd}, \tag{5.4}$$

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 Imbens 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 5.1 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.

5.5 Baseline results

5.5.1 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 5.4 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.

First-stage evidence from survey data and the complying subpopulation
The regression results of the first stage from Eq. 5.3 using NEPS data are shown for both the final sample and for certain subgroups in Table 5.3. The overall first-stage effect is very strong and is precisely estimated. To ease the interpretation of the compound instrument (defined in Eq. 5.1), 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 5.3: 2.08 (coefficient from the table) $\times$

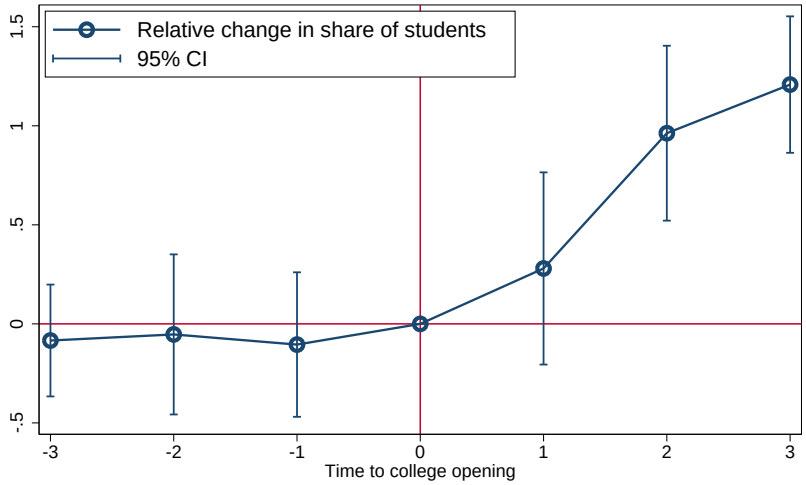


Figure 5.4: Relative change in the share of students in counties within 100km of college opening between 1962 to 1969

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, various issues, 1959–1991). The figure depicts the coefficients β_τ from the following "event-study" regression where β_0 is set to zero:

$$\begin{aligned}\ln(\#\text{students}_{bt}) \quad = \quad & \alpha_t + \sum_{\tau \in \{-7,-1\}} \beta_\tau \mathbb{1}\left[\max(t - t_b^{\text{opening}}, -3) = \max(\tau, -3)\right] \\ & + \sum_{\tau \in \{1,7\}} \beta_\tau \mathbb{1}\left[\min(t - t_b^{\text{opening}}, 3) = \min(\tau, 3)\right] + \gamma_b + \epsilon_{bt},\end{aligned}$$

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.

$K(0) \times {}^{15}/_{250} = 2.08 \times 0.4 \times 0.06 = 5$pp (rounded, see Eq. 5.1). 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

Table 5.3: First stage and some characteristics of complying mothers

	(1) Coefficient of the First Stage	(2) Share of the population	(3) Share of compliers	(4) 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.

is much stronger for women with lower-educated fathers as is evident from Table 5.3. 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 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.[14] 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.

5.5.2 The effect of college education on fertility

Starting with overall completed fertility, shown in panel A in column 1 of Table 5.4, 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 5.2, 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 5.4, 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

[14] 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 5.1. However, the degree of urbanization used here is only based on the number of inhabitants, not on the population density.

Table 5.4: 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; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

at the 10 percent level, the effect size is substantial. However, 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 5.5.1 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 5.4. 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.

5.6 Heterogeneity and potential mechanisms

5.6.1 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:[15]

$$\begin{aligned} d_{ig} &= \beta_0 + \beta_1 \widehat{D}_i + \sum_{g=17}^{40} \eta_g \mathbb{1}(age_{ig} = g) \\ &\quad + \sum_{g=17}^{40} \left[\gamma_g \mathbb{1}(age_{ig} = g) \times \widehat{D}_i \right] + \boldsymbol{X}_i' \boldsymbol{\beta}_2 + u_{ig}. \end{aligned} \tag{5.5}$$

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 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 5.5 shows the estimation results of Eq. 5.5 for the extensive margin of fertility in panel (a) and intensive margin in panel (b).[16] The bars state the baseline hazard rate of becoming a mother and the baseline probability of giving birth at a certain

[15] 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.

[16] 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

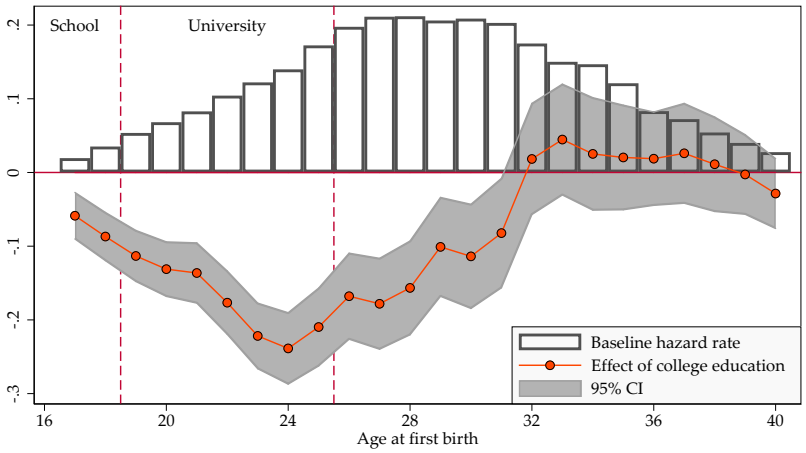
age in panel (a) and (b), respectively.[17] The oranges lines give the effect of college education on these baseline probabilities. For the sake of interpretation, we may think of the fertile ages as three phases for which we expect distinct effects: pre-college teenage years, years in college, and post-college 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 5.5. 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.

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 5.9 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.

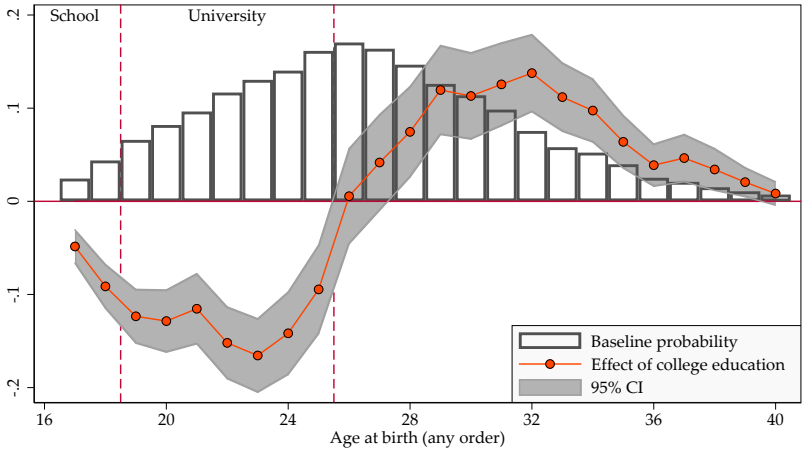
The third and final phase in fertile ages starts when individuals with a college education leave college – around the age of 25. At these ages college-educated women will reveal their preferences about fertility. Among the college-educated women who have not yet had a child, some may decide to remain 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

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.

[17] Note, the baseline rates plotted in Figure 5.5 state the unconditional means. On the contrary, η_g in Eq. 5.5 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 5.5: Timing of births

Notes: Both panels depict the age-specific regression coefficients from the second stage of the 2SLS model in Eq. 5.5 that capture the effect of college education. Panel 5.5a reports the effects of college education on the hazard rate of becoming a mother by age. Panel 5.5b depicts the respective effects on the probability of giving birth conditional on being a mother.

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 5.9). 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.

5.6.2 Opportunities and revealed preferences for career and family life

Table 5.5 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 5.5 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.[18] 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-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 5.6 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 5.1) could, if triggered by education, constitute an important mechanism as to

[18] 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.

Table 5.5: 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; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

why individuals put a stronger focus on family life or career opportunities. Columns 1 and 2 of Table 5.6 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 5.6 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, 2017). Column 5 of Table 5.6 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.

Table 5.6: 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; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.7 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. 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.

5.8 Appendix

Figures

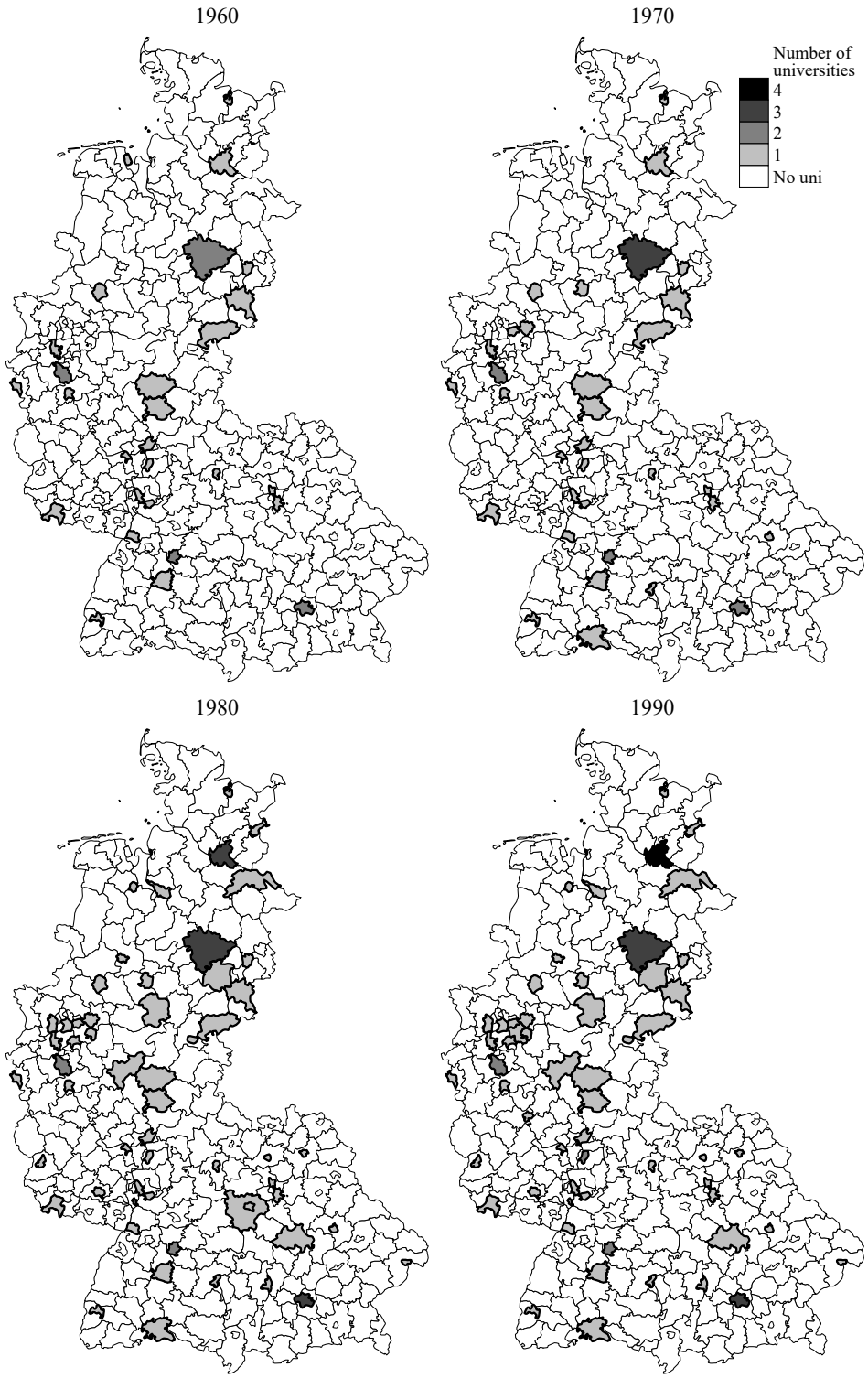


Figure 5.6: Spatial variation of colleges across districts and over time

Notes: Own illustration based on the German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, various issues, 1959–1991). 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.

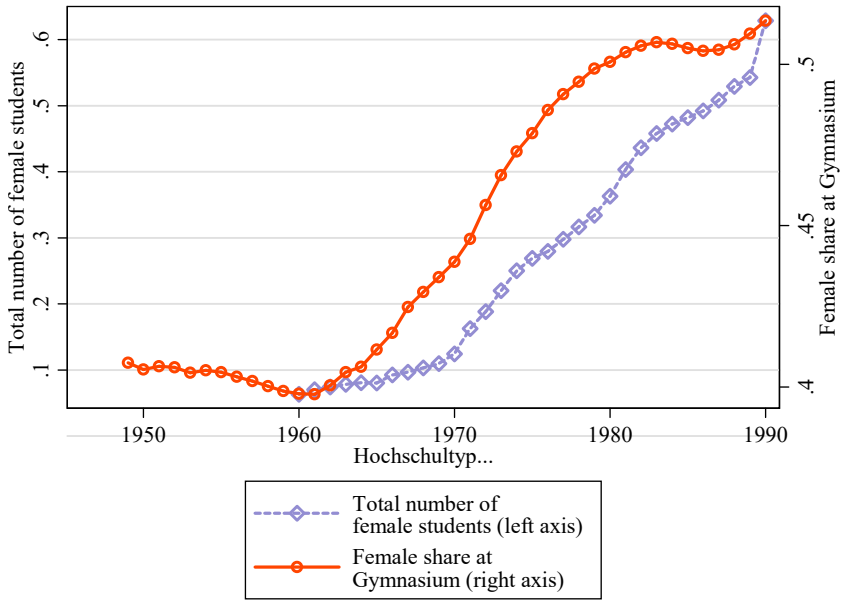


Figure 5.7: Trends in academic secondary school and college education for females

Notes: Own calculations using data from Köhler and Lundgreen (2015).

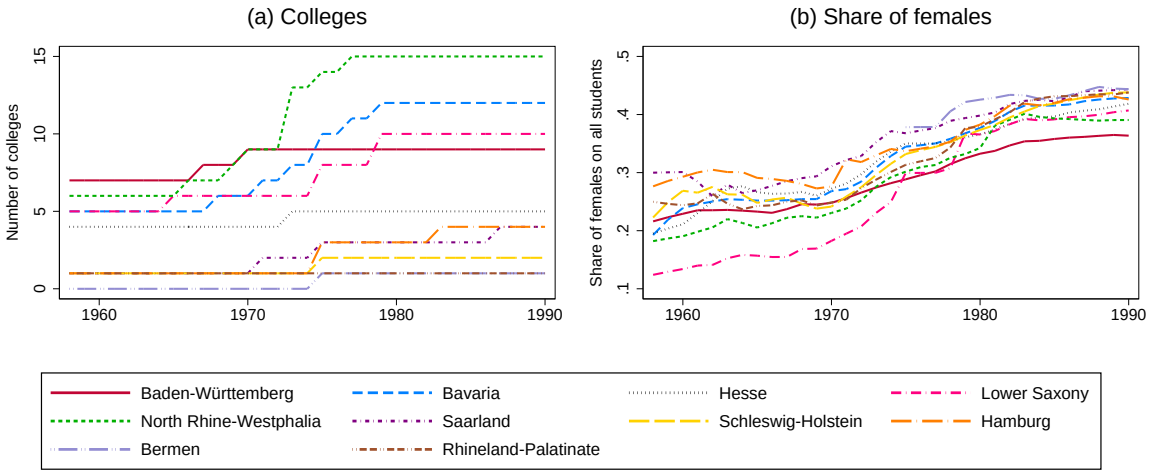


Figure 5.8: Trends in colleges and female students across federal states

Notes: Own calculations using data from the German Statistical Yearbooks 1959–1991 (German Federal Statistical Office, various issues, 1959–1991).

Tables

Table 5.7: 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
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

Continued on next page

Table 5.7 – continued

Variable	Definition	Respondents	
		with univ. degree	w/o univ. degree
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 5.8: 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, various issues, 1959–1991). Distances are calculated as the Euclidean distance between two respective district centroids.

Table 5.9: 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 5.5 and estimated according to Eq. 5.5. Unlike the figure, the baseline hazard and the baseline probability are stated by college status.

Chapter 6

More teachers, smarter students? Potential side effects of the German educational expansion[1]

6.1 Introduction

In recent years, the view that has ultimately prevailed is that education throughout the life course is important for acquiring skills that are decisive for, but not exclusively confined to, the labor market (Heckman et al., 2010; Chetty et al., 2011; Zimmerman, 2014; Kamhöfer et al., 2017). Teachers have a key role in creating environments and incentives for students to acquire these important skills, typically referred to in economics as the acquisition of human capital (Hanushek, 1971; Hanushek and Rivkin, 2006; Chetty et al., 2014a). Because of this key role, it is important to look at the leverage of educational policy on attracting high-quality teachers. If, for example, relatively less suitable individuals take up the teaching profession in response to changes of institutional arrangements, they could have a negative impact on the performance of their students. As all teachers educate generations of pupils over the course of their career, teachers can have a highly persistent impact on the skill acquisition of these pupils. Evidence from recent studies advocates such a persistent impact of teachers since resulting skill differentials at school may well spill over to later life by, for instance, affecting labor market performance (Chetty et al., 2014b).

In Germany, as in most industrialized societies, in the second half of the past century, educational policies were at the core of government institutional reforms. The goal was to increase access in particular to higher secondary education, namely the intermediate track (Realschule) and the academic track (Gymnasium), relative to the then-dominant basic track.[2] The quantitative expansion of both tracks was substantial even in relative terms: whereas only 20 percent of all pupils went to either one of

[1] This paper is published as: Westphal, M. (2017). More Teachers, Smarter Students? Potential Side Effects of the German Educational Expansion. Ruhr Economic Papers 721, RWI Essen. Funding by the Fritz-Thyssen Stiftung (grant no. 20.16.0.069) is gratefully acknowledged.

[2] At the same time, comprehensive schools (Gesamtschulen) were introduced. This school track, however, only played a minor role.

both tracks in the 1960s, this share had doubled by the end of the 1980s. This tremendous increase led to an upsurge in the demand for teachers.[3] Due to the educational expansion, roughly 150,000 new positions as teachers were created. These positions could not even theoretically be filled with basic track teachers, as these positions required more formal training.[4]

Did the implementation of this quantitative expansion lead to a diminishing quality of teachers? If, at any time, only the most motivated and able individuals took up the profession, an unanticipated and unprecedented increase in the demand for teachers could have encouraged less motivated and able individuals to eventually become teachers. The educational expansion is not only important because it created a demand-side variation in the labor market for teachers, it also captures a highly policy-relevant effect. Many of today's policies are often targeted at expanding public institutions like, for instance, the recent extension of the daycare sector and – potentially – of the future formal long-term care sector in Germany. These expansions exhibit characteristics that are similar to the educational expansion in the 1970s and '80s. Hence, knowledge about the past expansion is informative about how to efficiently implement new ones in the future.

The literature on teacher selection and its effects on student performance initially focused on identifying determinants of teacher selection. There is a large strand of literature that looks at the role of wage differentials between teachers' and the outside labor market (see, for instance, Britton and Propper, 2016, Loeb and Page, 2000, and Figlio, 1997, among others). Nagler et al. (2015) examine the consequences of business cycle-induced teacher selection on students' test scores. These studies find that a larger wage-differential leads to a diminishing teacher quality. Beyond wages, there are also further characteristics of the labor market of teachers subject to some studies. For instance, Lakdawalla (2001) determines the role of technological change and Bacolod (2007) considers the soared acceptance of female teachers. These studies likewise detect that teachers react to changed external incentives. Chetty et al. (2014b) go one step back by identifying the general impact of teachers on the human capital acquisition of their students. They uncover that replacing an average teacher with a teacher from the 5% quantile of the distribution of teacher quality raises the net present value of their lifetime earnings of the affected students by $250,000 per classroom.

I contribute to the literature on teacher selection and its effects on student performance mainly in two ways. First, this is the first study to specifically assess the consequences of one particular and major social change of the last 60 years – the educational expansion – not on those who are taught[5] but rather on those who teach. Insights into teachers are important since they are under a more direct control of policymakers who could then apply these insights to modifying the hiring process of teachers. Second, I am able to provide evidence on a much more homogeneous

[3] Because of a coinciding reduced student-teacher ratio, the demand for teachers was even higher than the increase in student numbers.

[4] In addition, the overall number of students in secondary education mechanically increased due to the changing track composition (academic track required four more years of schooling; the intermediate track, one year).

[5] Studies that focus on students comprise Siegler (2012) and Kamhöfer et al. (2017) for tertiary education, as well as Jürges et al. (2011) for secondary education.

group of high-skilled pupils who attend the academic track in Germany. This is in contrast to the existing literature that looked primarily at the comprehensive school system of the US or the UK.

To substantiate the exact specification of the educational expansion rate and the subsequent interpretation of the effects, I employ a simple theoretical framework of how marginal teachers affect the average quality of all teachers of a certain cohort. This model corroborates using relative changes in the stock of teachers in the federal state and year of the high school graduation as the educational expansion rate. This rate proxies the conditions of the teachers' labor market (and coinciding career incentives for those who are encouraged to become teachers). Subsequently, this proxy is related to the test scores of students instructed by a teacher decades later. By using these changes within German federal states that control and legislate the educational system within their borders, I am able to isolate the overall effects from a wide range of other effects. These confounding effects may arise because of unobserved third factors, for example, effects that go along with teachers' general experience or, more importantly, potential persistent differences in the quality of the educational system of the federal state. Concerning the former, for instance, the students' performance is measured decades later, long after the educational expansion was complete. Hence, I can disentangle the effects of the educational expansion that operates through teachers from the repercussions on students. Furthermore, I use a between-subjects difference-in-differences model to address the concern that good teachers may want to teach at good schools with better students. In the absence of any spillover effects, estimates of the cross-subject teacher environment on student test score relations (math teacher, reading scores and German teacher, math scores) identify confounding school selection effects, which can then be differenced out from the same-subject effects. If this teacher skill differential of educational expansion teachers is indeed driving the effect, I would expect this skill differential to also be reflected in some observed characteristics, such as subjectively assessed measures on intrinsic and extrinsic motivation.

To summarize the results, I find that students taught by teachers who witnessed an expanding teacher force in their federal state just after high school graduation score less in math and reading competence tests. By decomposing the effect into a component that is due to school selection (correlation between good teachers and initially good students) and a direct effect on test scores, I find that a significant share of the overall effect can be attributed to the direct effect of teachers on students. Teachers who graduated from high school in an average expansion year reduce the test scores of their students by 2 percent of an unconditional standard deviation (sd) relative to teachers that graduated in years with no expansion. The magnitude of the effect is comparable to related studies and is non-negligible. In providing an explanation for the identified test score differential, I look at the reported grade of the teachers' high school exit exam (Abitur) and examine further subjective measures of job choice and work ethic. I find that the educational expansion rate weakly predicts the academic achievement of teachers. In addition, educational expansion teachers are more extrinsically rather than intrinsically motivated.

The results have at least two important implications. First, as the policymaker certainly has more leverage in hiring good teachers than on directly influencing students or their family background, the conclusions of this paper are important for shaping

future policies. Connected to this, the second implication concerns today's and future expansions of public institutions in general, which become increasingly necessary in changing societies. Given the results of this paper, it seems crucial to not only invest in quantitative aspects, such as increasing the scope of arguably beneficial public institutions. Qualitative aspects are an important margin to invest in when implementing the expansion of these institutions. The substantial ongoing extension of daycare facilities (day nurseries and preschools) serves as a prime example. Since the educational expansion is paralleled by this expansion of daycare facilities, the results of this paper can rather easily be extrapolated to this setting.

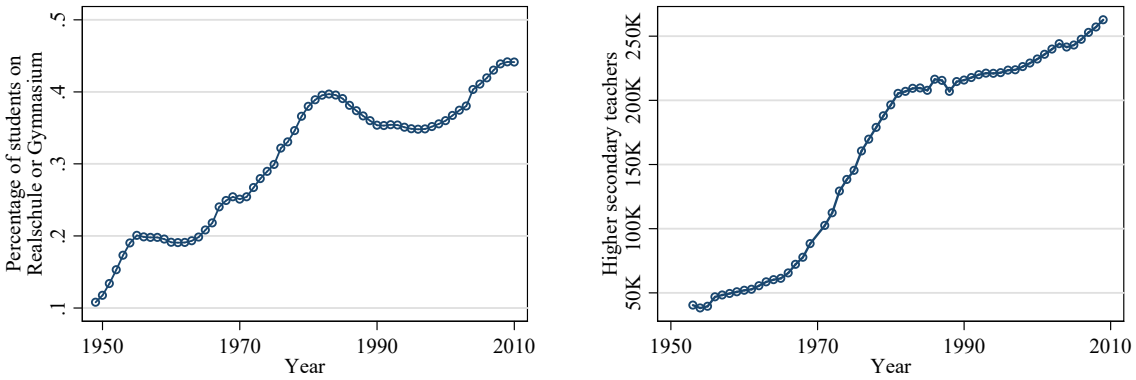
The remainder of the paper is structured as follows: Section 6.2 sets out the institutional background of the educational expansion in general and the teacher market in particular. Section 6.3 presents the empirical strategy that aims at estimating causal effects. Subsequently, a small theoretical mechanism is introduced that justifies the specification of the educational expansion rate and facilitates the interpretation of the results. Section 6.4 presents the data. Section 6.5 shows the main results of students' learning outcomes, assesses its robustness and presents supporting evidence on the characteristics of educational expansion teachers. Finally, Section 6.6 concludes.

6.2 The educational expansion and the market for teachers in Germany

In Germany, at least three things changed the notion of the scope of higher education, all of which took place roughly within 15 years. First, the view ultimately prevailed that education was key for social participation as a citizen, which served as a powerful intellectual and publicly influential argument to promote education (Dahrendorf, 1965). Second, as a consequence of its increased role internationally, reports of the OECD showed that Germany's system was internationally underdeveloped. This had, not least because of an influential book (Picht, 1965, which based on arguments set out in Picht, 1964), a huge impact on public opinion. The new and changed notion of education was reflected by the Social Democratic Party (SPD) making it the cornerstone of their new programmatic orientation: education policy was granted federal political importance by a party whose clientele traditionally came from educationally deprived strata (Osterroth and Schuster, 2000). Third, because of the Sputnik crisis in 1957, Western societies realized that they were trailing behind the Soviet Union. Opening higher secondary and tertiary education for a broader population was identified as being important for closing this gap in the long run. All these developments led to changes mainly in the supply of education, which shifted the composition of the students in terms of their field of study from public institutions traditionally being the most important employer of university graduates toward newly created jobs in engineering, administration, and the business sector (see, for example, Lundgreen and Schallmann, 2013).

The educational expansion also substantially affected secondary schools. This is visible in Figure 6.1a, where the share of pupils in the intermediate and academic track is plotted over time. The increased number of pupils required more teachers, also

because shifts in the track composition led to a mechanical increase in the average years of schooling (the intermediate track had one more year of schooling, the academic track four years more). Figure 6.1b illustrates the upsurge in teacher positions in higher secondary education over time: within 20 years, 150,000 additional teacher positions were created. The long-term repercussions of these new teachers are the subject of this paper. This requires looking at the dynamics that took place simultaneously, concerning, among others, teacher remuneration and the education of teachers in Germany. The current process of teacher training in Germany was imple-



(a) The percentage of intermediate and academic track students over time

(b) Dynamics in the stock of teachers on intermediate and academic track schools

Figure 6.1: Impact of the German educational expansion

Source: Köhler and Lundgreen (2015)

mented in 1917 for academic track teachers and was extended to include all teachers at primary and secondary schools up until 1970 (Köhler and Lundgreen, 2015). This process is called the "academization of the teaching profession" (Bölling, 1983; Köhler and Lundgreen, 2015). The training of all teachers from at least 1970 onward is set up as a two-stage process. All high school graduates with an academic track education (Abitur) are in principle eligible to be trained as teachers. Initially, teachers are educated at a university, commonly graduate in two specific subjects (Erstes Staatsexamen) and start a more education-specific vocational training at a certain school. After graduation from university – which takes 4.5 years – teachers graduate a second time (Zweites Staatsexamen) – which takes an additional 1-2 years – where teaching skills are tested. At the same time, there were also some changes in how teachers were remunerated. For example, one consequence of the academization was an increase in the salary level of teachers (Bölling, 1983). In addition, the teacher salary was leveled up to reduce the excess demand of teachers and to match their salary to wages in professions that required a similar qualification level. This, however, was largely completed before 1970 (Bölling, 1983; Köhler and Lundgreen, 2015) and therefore does not interfere with the study period (from 1970 onward).

6.3 Empirical strategy and theoretical mechanism

6.3.1 Empirical strategy

The aim is to compare "educational expansion" teachers (EETs) with teachers who were not influenced by the educational expansion. I consider EETs as being individuals who started their teacher training and education during the massive demand increase that occurred during the educational expansion. On average, these teachers may differ because of some marginal teachers. These marginal teachers are a subset of all EETs and only took up the teaching profession because of changed career incentives (Ashraf et al., 2014). For instance, an awareness of the possibility of eventually becoming a teacher may have surged. If the educational expansion occurred in certain years and not in others, I could simply compare EETs with teachers who started their education after or before the educational expansion. The time scale at the bottom of Figure 6.2 illustrates this hypothetical clear temporal demarcation. However, the time of the educational expansion cannot be clearly defined. Yet, it can be exploited that the federal states in Germany have discretion over when, where, and to which extent to increase the capacity of the (secondary) educational system. Additionally, federal states decide on the curriculum in schools and in teacher training. Because of this institutional peculiarity, the mobility of teachers between federal states is low (Table 6.8 shows that nearly three quarters of teachers stay in the federal state where they graduated from high school). Consequently, I use the relative expansion of the teacher force at the federal state level to capture the part of the educational expansion that affected the job prospects of future teachers.

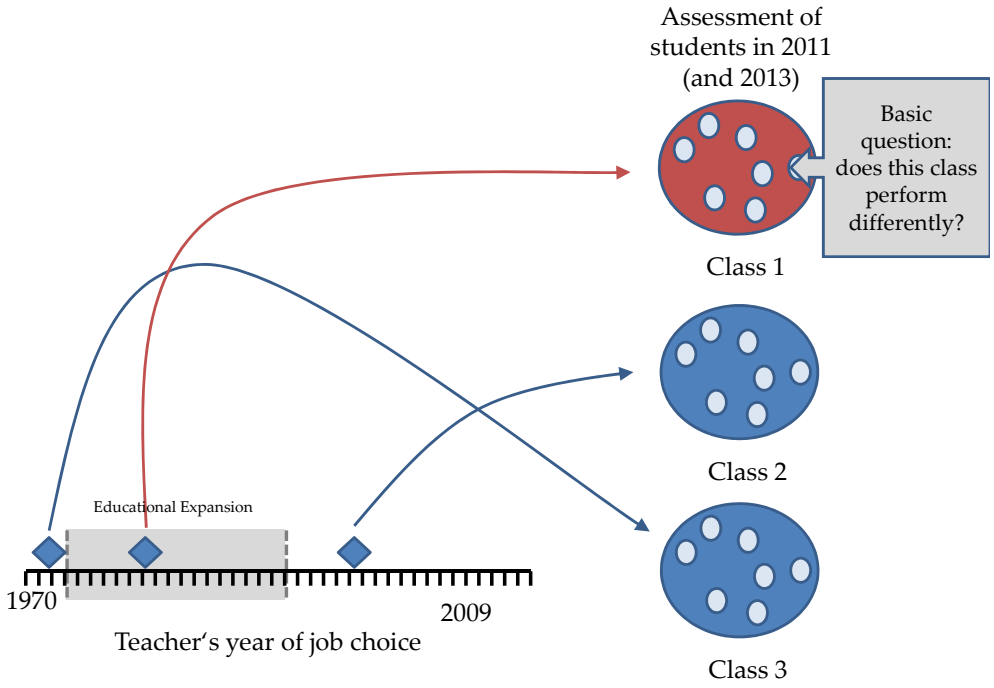


Figure 6.2: Illustration of the fixed effects setup

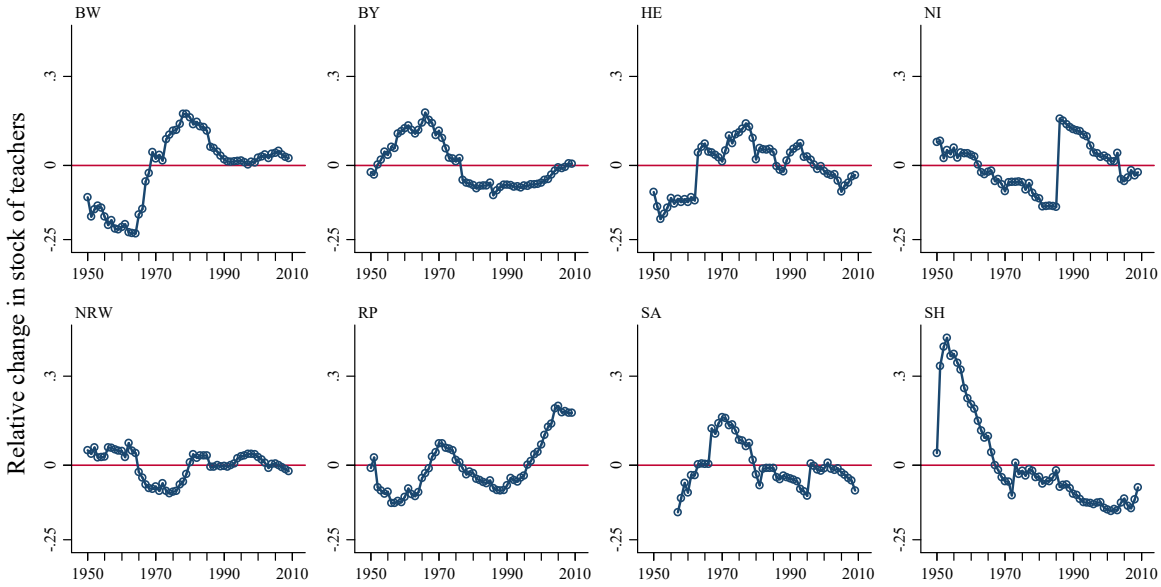


Figure 6.3: Relative changes in the stock of teachers by non-urban federal states over time

Notes: the time-series are residuals from a population-weighted regression of the stock of teachers on federal state and year fixed-effects. The data are based on administrative records and taken from Köhler and Lundgreen (2015).

This relative change in the stock of teachers over time and by federal state is depicted in Figure 6.3. In this graph, the differences in the timing as well as in the intensity with which the educational expansion was carried out are clearly visible. Each panel in Figure 6.3 depicts the West-German 'Flächenländer' (the urban federal states Berlin, Bremen, and Hamburg are excluded for the sake of clarity). The graph illustrates the different developments in the teacher market. If the teacher force of any given federal state grows faster relative to all federal states in a given year and faster than the own average growth rate, the relative changes plotted in Figure 6.3 are positive. Conversely, if the growth of the teacher force is lower than the trend in the federal states as well as the overall yearly change on the federal level, the relative change is negative. Another way to interpret the relative change in Figure 6.3 is by relating the number of (marginal) EETs to the number of teachers that were projected to be needed in the absence of the educational expansion, which is clarified in the next subsection.

Figure 6.2 also illustrates the general data structure that is exploited in the empirical approach. The three classes on the right-hand side of Figure 6.2 represent all fifth and ninth grades in the data. The pupils in those classes are subjected to objective tests on their math and reading performance. These test scores can be linked to teachers that teach the respective subject: German teachers are assigned to reading test scores and math teachers to math test scores. The educational expansion rate is merged to those teachers based on the federal state and the year (birth year plus 19) of their high school graduation. The effect of the educational expansion on students' learning outcomes may then be picked up by β_{FE} in the following regression:

$$y_{i_{\tau l} j_{st}} = \beta_{\text{FE}} \ln(\#\text{teachers}_{st}) + \theta_s + \pi_t + \eta_l + \mu_\tau + \mathbf{X}'\boldsymbol{\rho} + \epsilon_{i_{\tau l} j_{st}} \tag{6.1}$$

where y measures the test scores of student i in year τ taught by teacher j at a school in state l who received his secondary school diploma in state s in year t. Because of the twofold fixed effects at the teacher level (θ_s and π_t), β_{FE} is essentially identified by relative deviations from the state-specific mean and the average yearly change across all federal states.[6] These deviations are exactly what is depicted in Figure 6.3. In addition to the teacher level fixed effects, student level fixed effects are also employed (η_l and μ_τ) to control for any persistent differences between the years and the federal states of the schools. Moreover, $\mathbf{X}$ may contain further covariates to possibly control for class composition, depending on the exact specification. In this fixed effects model, β_{FE} may pick up the effect of teacher quality on students' learning outcomes, if changes $\ln(\#\text{teachers}_{st})$ only capture the difference in teaching quality between EET and non-EET (see next subsection) with all else held fixed. However, one could still be concerned that skilled teachers have better opportunities to choose the school they teach in. Such a selection would confound β_{FE}.

To break the correlation between the initial skills of the students and teacher quality, variation between subjects (math and German) is exploited. Table 6.1 shows how this information helps to improve the identification. As every student has a German

Table 6.1: Setup of the difference-in-differences approach

	Math Scores	Reading Scores
Math Teacher	Treatment ($D = 1$)	Control ($D = 0$)
German Teacher	Control ($D = 0$)	Treatment ($D = 1$)

and a math teacher and is assessed in both reading and math skills, there are four possibilities for using the test score observations of a certain student (indicated by the gray-shaded cells). First, the math score is evaluated with respect to the exposure to the educational expansion (the relative changes depicted in Figure 6.3) of his math teacher. Second, reading scores and the exposure of the German teacher can be used. Both assessments are reflected in β_{FE}. This coefficient captures the direct effect of teacher quality plus, potentially, some school sorting effect. Moreover, also assessing across subjects can be informative: relating math scores to German teachers and reading scores to math teachers. Estimating Eq. (6.1) using this cross-subject test score-teacher association yields the school sorting effect and potentially also the same spillover effect. In the absence of a spillover effect, the school sorting effect

[6] Thus, it can also be termed a difference-in-differences model with continuous treatment. The reason why I refer to this model as 'fixed effects' is to clearly separate the wording from the difference-in-differences model that is employed later on.

is identified and can be substracted out of β_{FE}. This can be directly done by defining a treatment and a control group (indicated by the treatment variable D) and by estimating the following model:

$$\begin{aligned} y_{i_{\tau lF} j_{tsF}} &= \alpha + \beta_{\text{DiD}} \ln(\#\text{teachers}_{st}) \times D + \delta \ln(\#\text{teachers}_{st}) + \gamma D \\ &\quad + \theta_{t\times D} + \theta_{s\times D} + \eta_l + \mu_\tau + \mathbf{X}'\boldsymbol{\rho} + \epsilon_{i_{\tau lF} j_{tsF}} \end{aligned} \tag{6.2}$$

Because this model differences out the school sorting effect, it is a difference-in-differences approach (DiD). The treatment group comprises students' test scores and teachers from the same subject and is indicated by the treatment indicator D taking the value 1. The control group, on the other hand, connects students' test scores and teachers between the subjects (math and German). This relation is indicated by $D = 0$. To facilitate interpretation, the fixed effects of the state and the year of the teacher's high school graduation are now interacted with D.[7] Finally, standard errors for β_{FE} and β_{DiD} are clustered on the federal state and year level of the teachers' high school exit exam since this is the level where the hiring of teachers occurs.

Besides a school sorting effect, this regression automatically purges all individual and also class and school fixed effects. If the assignment of German and math teachers to classes is mean-independent of teacher quality and of the relative, subject-specific skills of the class, the coefficient β_{DiD} identifies the causal effect (see Appendix 6.7 for a clear list of the identifying assumptions). Also, in the case of spillover effects, the school sorting effect is differenced out. Then β_{DiD} is a lower bound for the gross effect of teacher quality, since school sorting and spillover effects are both likely to be positive. However, the literature only finds weak evidence of the existence of spillover effects (Koedel, 2009). In robustness checks, however, I will scrutinize these spillover effects directly.

6.3.2 Theoretical mechanism

In response to the educational expansion, different individuals could have been encouraged to become teachers who also exhibit different career incentives. Why is that? As in every market, the labor market for teachers can also be characterized by two major forces, demand and supply. Regarding the former, the federal state s

[7] In the difference-in-differences equations as in (6.2) interpreting β_{DiD} as being identified from deviations from state and year-specific means would not work. To get these deviations, regress $\ln(\#\text{teachers}_{st}) \times D$ on the respective fixed effects (by the Frisch-Waugh-Lovell Theorem, a 'second stage' regression of y on ω_{st} and $\ln(\#\text{teachers}_{st})$ would yield the same coefficients as in Eq. (6.2) without interacted fixed effects):

$$\begin{aligned} \ln(\#\text{teachers}_{st}) \times D &= \mu_t + \eta_s + \omega_{st} \\ E[\ln(\#\text{teachers}_{st}) \times D] &= \delta_t \times \Pr(D) + \pi_s \times \Pr(D) + \epsilon_{st} \times \Pr(D) \end{aligned}$$

Applying the law of iterated expectations shows that the essential variation that identifies β_{DiD} is deflated by $\Pr(D)$. Using D-specific fixed effects adjusts for this deflation directly. Hence, interacted fixed effects are necessary in order to interpret β_{DiD} as deviations from the state-specific as well as the year-specific mean.

may project the demand for teachers in year t based on the expected number of academic track pupils, $\mathbb{E}_{st}P_{st}$. Also, the fraction of the teaching force that retires, δT_{st} may contribute to the demand for new teachers. In total, the overall demand for teachers can be expressed as $D_{st}(\mathbb{E}_{st}P_{st}, \delta T_{st})$. Because the federal states hire based only on how many students are enrolled or will enroll into the secondary educational system, supply-induced demand is unlikely to occur. Therefore, the demand can be seen as independent of the potential quality of teachers. It is exogenous to potential teachers.

Supply, on the other hand, is determined by the number of academic track graduates in year j and federal state s, as the job mobility between federal states is rather limited. Each individual within a cohort and a federal state has a net benefit of teaching $B(j_{st})$. This net benefit is the benefit of working as a teacher minus the benefit of working in the next best occupation. Hence, having the highest net benefit does not necessarily mean being the best teacher. It means that the skills or preferences of this individual are most teacher-specific. This benefit may depend on a vector of individual characteristics $\boldsymbol{S}_{j_{st}}$ of the potential teacher j_{st} that can be closely related to teacher quality $Q_{j_{st}}$. For instance, this vector may comprise intrinsic motivation to teach, specific teacher quality, and general skills among others. Thus, individuals with the highest benefit are most likely to be intrinsically motivated and have a high teacher quality. Similar to a Roy-type selection model of occupational choice (Roy, 1951), individuals will start teacher training based on this net benefit. But for individuals at the margin of becoming teachers, the decision may additionally depend on external market forces, such as the recruiting policy of the federal state. These individuals are less determined to join the teaching profession. Hence, extrinsic factors such as chances of eventually being hired as teachers, the prestige of the teaching job, or the relative salary are more important to those individuals.

Figure 6.4 plots the supply and demand forces. On the horizontal axis the share of academic track graduates in year t and federal state s with at most a certain teacher net benefit is depicted (for clarity, the scales are exaggerated). This share is mapped on the net benefit of being a teacher for all individuals in this cohort. Along the horizontal axis, the net benefit decreases. Thus, this supply function is equivalent to the quantile function of individuals having at most a certain net benefit. This is also called the inverted complementary distribution of the teacher net benefit: $q_{j_{st}} = (1 - F(B_{j_{st}}))^{-1}$.

In the absence of the educational expansion – which is targeted at increasing the share of each birth cohort with an academic track education – a fraction p_1 of each birth cohort can become teachers. This fraction depends on the demand for teachers D_{st}, which introduces external equilibrium factors to influence individual choices. Most likely, the individuals who become teachers are among those with the highest net benefit and implicitly exhibit those characteristics $\boldsymbol{S}_{j_{st}}$ that are better suited for being a good teacher. Note that D_{st} can also monotonously change from year to year in response to a constant fraction of teachers retiring or because the cohorts of students who transition to academic track education and those of high school graduates are constantly growing in the federal state.

In response to the educational expansion, there is an exogenous increase in D_{st}, denoted by ΔD_{st}. This has two notable consequences that outline the tradeoff between the quantity and quality of teachers. First, an additional fraction p_2 (the marginal

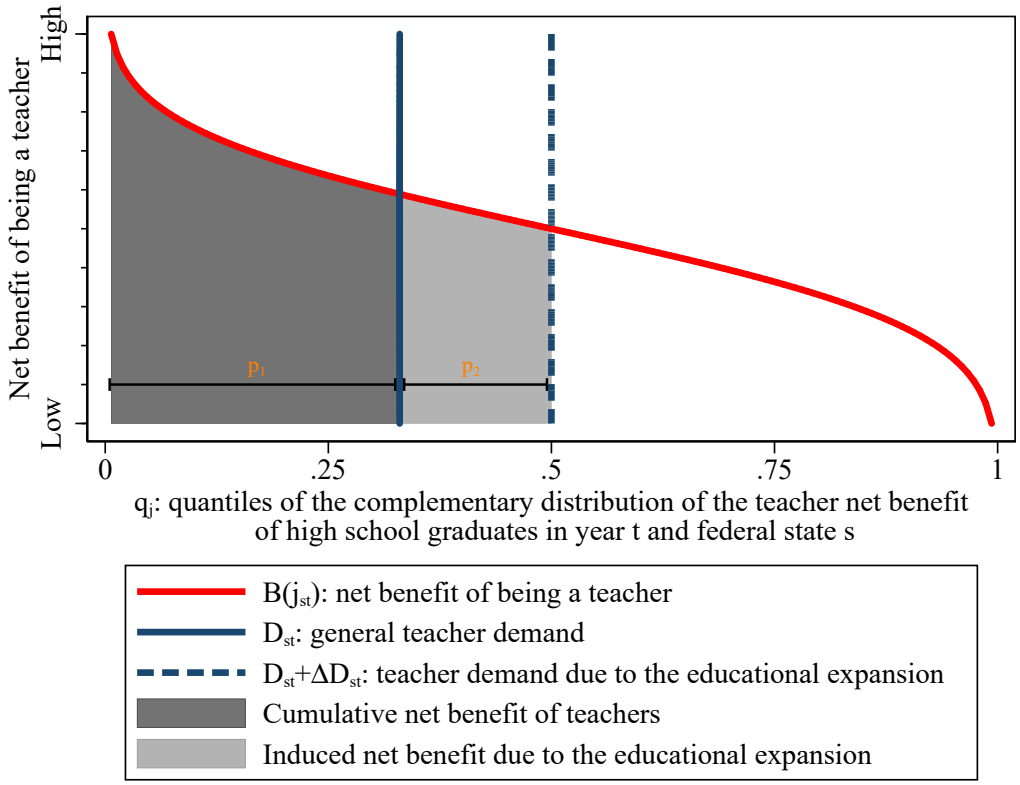


Figure 6.4: Possible impact of the educational expansion on the job market for teachers

teachers) of the high school graduate cohort that witnesses the demand increase for teachers in year t and federal state s decide to become teachers (the share of EETs amounts to $p_1 + p_2$ if $p_2 > 0$). The second consequence is that the average net benefit of all teachers – and therefore, most likely also the corresponding teacher quality – diminishes. In this model, the average net benefit of the p_1 teachers from a high school cohort in a federal state in normal years amounts to $\overline{B\left(D_{st}\right)} = \int_0^{p_1} B(j_{st})dF(q_{j_{st}})$ (depicted by the dark gray area in Figure 6.4) where $F(q_{j_{st}})$ is a uniform distribution (quantiles of a population are uniformly distributed). Accordingly, the average net benefit of those individuals who become teachers due to the educational expansion (marginal teachers) is: $\overline{B\left(\Delta D_{st}\right)} = \int_{p_1}^{p_1+p_2} B(j_{st})dF(q_{j_{st}})$ (indicated by the light gray area). The overall average net benefit (light and dark gray-shaded areas) of a teacher

cohort t in federal state s that witnesses a teacher expansion (or contraction, $p_2 \neq 0$) can then be expressed as:

$$\underbrace{\overline{B\left(D_{st}+\Delta D_{st}\right)}}_{\text{Average net benefit of EETs}} = \overbrace{\overline{B(D_{st})}}^{\substack{\text{Average net}\\ \text{benefit}\\ \text{of non-marginal}\\ \text{EETs}}} + \underbrace{\frac{p_2}{p_1+p_2}}_{\substack{\text{Fraction of}\\ \text{marginal EETs}\\ \text{to all EETs}}} \overbrace{\left[\overline{B(\Delta D_{st})}-\overline{B(D_{st})}\right]}^{\substack{\text{Net benefit differential}\\ \text{between marginal and}\\ \text{non-marginal EEsT}}} \tag{6.3}$$

This expression explicitly shows how the average individual net benefit changes with respect to newly entering marginal teachers. The same effect applies not only to the net benefit but also to teacher quality if the benefit is monotonously related to the ability to teach (which is likely): $\overline{Q\left(D_{st}+\Delta D_{st}\right)}$. This equation is important in mainly two respects. First, $p_2/(p_1+p_2)$ is similar to the employed educational expansion rate as depicted in Figure 6.3. This rate is p_2/p_1. In the appendix, I show that the empirical results are insensitive to employing p_2/p_1, or $p_2/(p_1+p_2)$. Thus, it shows that the effect of the educational expansion on the labor market for teachers can be measured by the relative share of incoming teachers (rather than, for instance, the absolute number of teachers). As this is achieved by the log-specification, Eq. (6.3) justifies its use as the preferred specification in the empirical models of Eq. (6.1) and (6.2). Using $\ln(\#\text{teachers}_{st})$ mechanically adjusts the effect from all EETs (p_1+p_2) to the marginal teachers (p_2, as a local average treatment effect adjusts the effect to the complying population) – the EET (light gray area) – and thus does not average the effect over all teachers in a particular cohort (light and dark gray areas). In this sense, one can think of this approach as also being an instrumental variables approach.

The second reason for why Eq. (6.3) is useful is for interpreting the results later. As outlined in the empirical strategy, I test whether p_2/p_1 is correlated with the test scores of students. If it is correlated, the effects in β_{FE} and β_{DiD} are given for the average change in teacher quality $[\overline{Q(\Delta D_{st})}-\overline{Q(D_{st})}]$ (if this is the exclusive driver behind the effect on student performance) averaged over all years and federal states (changes in $\overline{B(D_{st})}$ are captured by the fixed effects, π_t and θ_s in the regression models (6.1) and (6.2)). If this quality differential was observed, one could regress $[\overline{Q(\Delta D_{st})}-\overline{Q(D_{st})}]$ in a first state on $p_2/(p_1+p_2)$. Then, the reduced-form effect can be adjusted not only to the marginal teachers but also to a one-unit increase in teacher quality. These two features imply that the effects of the educational expansion can be precisely identified. In contrast, the effect of latent teacher quality on students' learning outcomes is a reduced-form effect (in terms of teacher quality) as teacher quality is unobserved.

6.4 Data

6.4.1 Sample selection and student-teacher linking

This study exploits the National Educational Panel Study (Blossfeld et al., 2011b). The NEPS has a multi-cohort design and covers the educational trajectories of all individuals from six different stages of life. Specifically, I use the third (SC3) and the fourth (SC4) starting cohorts. SC3 comprises individuals that attended the fifth grade, whereas SC4 contains individuals from the ninth grade at the start of the school year 2010/2011. Compared to any survey data in Germany, the advantage of the NEPS is that it includes information on both the students and their teachers. The design of the questionnaire is equivalent across both cohorts. Hence, individuals and teachers from both starting cohorts can be pooled together in one sample.

The sampling population are all German fifth and ninth graders in 2010. In a first step, 234 schools are sampled (Skopek et al., 2012). All students in grades 5 and 9 from these schools are asked to participate in the survey. Since the NEPS is a panel survey, it follows these students as they move through the education system, including general education and occupational training. The survey also extends to the students' parents and the teachers in math, German as well as the class teachers. Teachers are interviewed once and can be linked to the respective class they teach. Information on teachers include year of birth, their high school graduation, their college education, retrospective determinants of their occupational choice and their attitude toward their job as a teacher.

Several restrictions need to be imposed on the data. From initially 1,206 teachers and 9,042 students that attend higher secondary schools in West Germany (the educational expansion did not take place east of the Iron Curtain, including East Germany), I restrict the sample to academic track schools. This group of students is high-skilled and mostly homogeneous in their abilities. Furthermore, I keep only teachers who either teach math and German (thereby dropping the class teachers). Both restrictions reduce the sample to 345 teachers and 4,259 students. Lastly, I restrict teachers to being younger than 60 years old as older teachers might already anticipating retirement . Therefore, I additionally drop 23 teachers. This means the oldest teachers in my final sample made the decision in 1970 to become a teacher, which is after the adjustment processes of teacher salaries and teacher training had finished.

Figure 6.5 shows the number of teachers in my sample by subject over time. There is approximately an equal amount of math and German teachers, and only a negligible minority teach both subjects. As is visible by the co-movement in the number of subject teachers over time, there is more variation over time than between subjects. In the NEPS teacher force, there are many teachers who graduated from high school (at age 19) in the 1970s. The 1980s are characterized by a saturated teacher force and relatively fewer hirings, which is also reflected in Figure 6.5. In the 1990s and 2000s (until 2005) the number of teachers in the sample increases again.

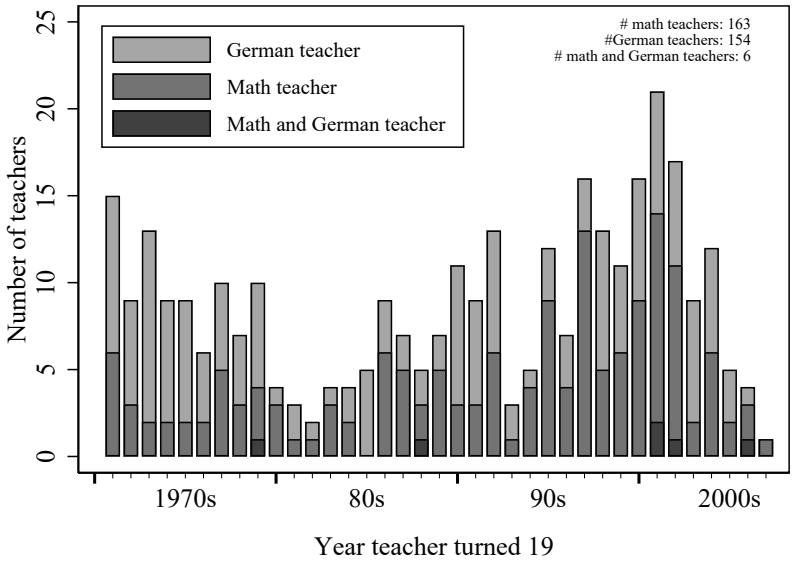


Figure 6.5: Number of teachers by year and subject
Notes: own calculations using NEPS data.

6.4.2 Test score data

THe unidimensional competence scores serve as the main outcome variables in reading and mathematics.[8] These scores have been assessed in tests conducted between November and January of a school year. As the school year usually starts in August, teachers can impact the test scores of their students through lessons in the first three to five months of the school year (on average, it is 3.72 months). Teachers cannot control the results of the test as these are conducted by trained NEPS interviewers. The scores are assessed by multiple choice questionnaires that every pupil has to fill in. The answers to these questions are aggregated by a weighted maximum likelihood estimation (WLE, Pohl and Carstensen, 2012). WLEs in the first wave are constrained to having a mean of zero. Values above zero therefore indicate abilities above average. This makes the scores comparable across the waves and cohorts. The variance of the WLE scores is not restricted.

Math competence score
Mathematical competence is targeted at measuring the "ability to flexibly use and apply mathematics in realistic situations" (Schnittjer and Duchhardt, 2015, p.2). Mathematical competence is assessed by 24 items in grade 5 and 22 items in grade nine on several domains.[9] For both grades, the test is designed to take 28 minutes in total. Examples of multiple choice questions include the following: "Mr. Brown owns a rectangular plot, which he wants to fence. After some calculations he buys 40m fence. The plot is 8m wide. How long is the plot?"

[8]The data are cleaned from effects of position and order. This is achieved through a random assignment of the order of the two tests to respondents (Durchhardt and Gerdes, 2012).

[9]Quantity is captured by eight items, space and shape in total have five, change and relationships six and Data and chance five.

Reading scores
Understanding and using written texts is an important skill and a prerequisite for participating in cultural and social life (Gehrer et al., 2012). The reading score test is designed to measure those skills. As German lessons are designed to let students acquire the exact same skill, the reading score skills can be attributed to the domain of the German teacher. In order to accurately assess these skills, it is distinguished between five "text functions and associated text types" (informational texts, commenting or argumenting texts, literary texts, instructional texts, and advertising texts). Within the time of the test (also 28 minutes), the test participants are given the five types of texts ranging from informational to literary texts. Each type of text is associated with a different skill. Texts are adjusted to the lexical level, difficulty, and thematic orientation of the specific cohort and age level. The participants are ask to read a short text, before answering multiple choice questions Right after having read each text, .

6.4.3 Descriptive statistics

Table 6.2 presents some descriptive statistics with respect to the educational expansion status of the teacher. For the sake of simplicity, the educational expansion rate p_2/p_1 is discretized at a threshold of zero. According to this definition, 2,203 students are taught by EETs, while 2,816 students have a non-EET.

Educational expansion teachers teach students with a worse test score (0.92 vs. 0.94 for reading and 1.02 vs. 1.08 for math) – a first descriptive indication of an effect. The gender of the students is balanced between EET and non-EET. German teachers are less likely to be classified as EET according to my definition. Potentially, this is because math teachers possess skills that make them react more sensitively to the changed career incentives of the educational expansion. The next four characteristics refer to the educational expansion rate. It is shown as its raw values (the log stock of teachers) and as the effective variation (demeaned by year and federal state fixed effects). These measures are presented separately for math and German teachers.

The average class size differs slightly between EET and non-EET (17.4 vs. 18.4). The instructional time (time from start of the school year to the assessment of the test score) also varies slightly according to the educational expansion status of the teacher. In the overall sample, slightly more students are in the initial ninth grade (SC4). The students in this grade have a higher chance of being taught by an EET. Within the initial fifth grade (SC3), 56 percent of the observations come from the second wave (all observations from SC4 are assessed in the first wave). This statistic also varies somewhat by the educational expansion status of the teacher. Although the sample appears to be slightly imbalanced in these respects, the empirical strategy and the robustness checks rule out that imbalances between cohorts and waves can carry over to the identification of the main effects.

Table 6.2: Descriptive statistics

	Educational expansion teachers: $p_2/p1 > 0$		Non-expansion teachers: $p_2/p1 \leq 0$	
	Mean	sd	Mean	sd
Test scores				
Reading	0.92	(1.13)	0.94	(1.14)
Math	1.01	(1.12)	1.08	(1.14)
Student characteristics				
Share female pupils	0.53	(0.50)	0.52	(0.50)
Teacher characteristics				
Share German teachers	0.49	(0.50)	0.55	(0.50)
Treatment, the relative expansion in the stock of teachers:				
Raw values:				
$\ln(\#\text{teachers}_{st})$ for German teachers	10.30	(0.61)	10.22	(0.49)
$\ln(\#\text{teachers}_{st})$ for math teachers	10.30	(0.80)	10.24	(0.60)
Effective variation: $p_2/p1$ (plotted in Figure 6.3):				
$p_2/p1$ for German teachers	0.05	(0.04)	−0.03	(0.02)
$p_2/p1$ for math teachers	0.03	(0.03)	−0.03	(0.02)
Class characteristics				
Class size	17.43	(5.94)	18.42	(5.35)
Minimum instructional time of teachers	3.58	(0.57)	3.86	(0.62)
General characteristics				
Share from SC4	0.64	(0.48)	0.53	(0.50)
Share from second wave among SC3 observations	0.60	(0.49)	0.53	(0.50)
Number of student-teacher-course-wave observations	2,203		2,816	

Notes: This is the effective variation, which refers to the variation in $\ln(\#teachers_{st})$ when all other variables, most importantly federal state and year fixed effects, are held fixed: the residual of log stock on year and federal state fixed effects, which are relative changes in the federal state-specific stock of teachers from the general expansion trend across all federal states.

6.5 Results

6.5.1 Effects on students' learning outcomes

Table 6.3 presents the estimation results from Eq. (6.1), the baseline fixed effects results specification by subject. It is a first step in clarifying whether individuals were encouraged to become teachers by the educational expansion and are now teaching students that today perform differently at school. The first line of Table 6.3 shows

Table 6.3: Fixed effects results for math and reading competence

	Math teacher – math competence			German teacher –reading competence		
	(1)	(2)	(3)	(4)	(5)	(6)
$\ln(\#\text{teacher}_{st})$	-1.237^{**}	-1.295^{**}	-1.382^{***}	-0.382	-0.650	-0.743
	(0.536)	(0.510)	(0.509)	(0.500)	(0.456)	(0.464)
Further condition on:						
_ Cross-subject competence score		✓	✓		✓	✓
_ Federal state of school FE			✓			✓
Observations	2,713	2,620	2,620	2,437	2,399	2,399
Number of teachers	168	168	168	158	158	158

Federal-state-by-year-level clustered standard errors in parentheses, $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$. All columns refer to a separate regression with additional federal state and year fixed effects plus all effects indicated.

the association between the change in the stock of teachers in the year and the federal state of high school graduation and the respective test score of the pupils that they taught in the survey year. The first three columns refer to math teachers and the associated math score of their pupils, the last three columns are results for German teachers and the reading score of their pupils. On average, the math competence score is 0.0127 points lower for every 1 percent that the stock of teachers increased relative to the overall trend in the year the teacher turned 19 and decided on his future job (as reflected by p_2/p_1). Two things are worth noting: first, the result is non-negligible in magnitude and suggests that teachers play an important role. Why is the coefficient plausible? The mean effective variation that identifies β_{FE} (the mean absolute deviation of the residual of a regression of $\ln(\#\text{teachers}_{st})$ on all the controls) shows that the mean change in the stock of teachers was 4.33 percent on average. Multiplying β_{FE} with this variation and dividing by the standard deviation in math competence indicates that 5.3 percent[10] of a math score standard deviation can on average be attributed to the educational expansion (if interpreted as causal). As I will try to demonstrate below, this magnitude fits well into what previous studies found. The second notable point is that the effect is robust toward the inclusion of important control variables that may mitigate the role of school selection: including reading competence is supposed to capture the general ability of the student whereas state of school fixed effects should control for persistent migration patterns of teachers within Germany. Because the results are robust toward the inclusion of these fixed effects, migration of teachers (shown in Table 6.8) does not affects the results.

For reading competence, the results are somewhat different, although the direction of the effect is unchanged. Having a teacher that was gradually exposed to a higher degree of the educational expansion – as measured by a 1 percent increase in the relative change in the stock of teachers – goes along with having a 0.0038–0.0074 lower

[10] Calculation: 1.3827[coefficient] $\times$ 4.33[[mean absolute deviation in %]/1.13[sd of test score].

score in reading competence depending on the specification. Applying the same calculation as above yields the fraction of a standard deviation in reading scores that can be attributed to the educational expansion (again, a causal interpretation) shows that this fraction amounts to 2.79 percent. Note, however, that none of these results are significant at the 5 percent level. Moreover, recall that the finding of smaller effects on reading competence is in line with the literature where, for instance, Nagler et al. (2015) also find smaller effects of recession teachers on the reading value-added measure of their students. Also, Chetty et al. (2014a) report a smaller value-added transmission on reading compared to math scores. In the context of this paper, this finding can be due to two reasons. First, German teachers may generally have a lower leverage on reading scores whereas the math score might better capture what is taught in the lessons. Second, the German teachers might have reacted differently to the educational expansion such that the effect on teacher quality is not that pronounced. One reason for this can be the potentially better outside option for math teachers.

How likely is it that these effects are attributable to the teacher and not to some unobserved class, school, or individual characteristics? To answering this important question, I now turn to the difference-in-differences estimation outlined in equation (6.2). Its results are presented in Table 6.4. These are the main results of the paper, since it comes closest to answer the question – what is the effect of teacher selection induced by the educational expansion on the learning outcomes of today's pupils. To approach an answer, I first pool data from all the cells of Table 6.1 into one comprehensive sample. As a result, I have one pupil by test score observation by teacher (see an example data set in Table 6.9), but every pupil can now appear in the sample up to four times. This approach allows me to use information on all teachers and students simultaneously. To adjust standard errors to this restructuring, standard errors remain clustered on federal state by year level as before and throughout the whole analysis. In Table 6.4 the main coefficients are presented, with subsequently added control variables as one moves from the left to the right columns. The main and most important effect listed in the first line ($\ln(\#\text{teachers}_{st}) \times D$). It captures the additional effect of the educational expansion of teachers that teach the corresponding subject measured by the outcome variable (math competence for math teachers and reading competence for reading teachers). This effect is significant and robust toward the inclusion of further fixed effects (columns 2–8): explicit subject fixed effects do not change the result (column 2; as they are implicitly incorporated in Eq. (6.2)), neither do the characteristics of the teachers (column 3). Including state of school fixed effect slightly inflate the effect (column 4), whereas cohort, wave, class, test month fixed effects nor even state-specific trends impact the coefficient any further. Causally interpreting this effect means: every 1 percent of a higher relative demand for teachers would attract teachers that – on average – reduce the subject-specific test scores of their pupils by 0.00822 to 0.00966. Conducting the same exercise as above and taking the mean effective variation that identifies the effect for $\ln(\#\text{teachers}_{st}) \times D$ – which in this setting amounts to 2.38 percent – shows that 2.02 percent of the overall standard deviation can on average be attributed to the educational expansion.[11] The difference between this fraction and the average effects of the FE model in math and reading (5.3 for math and 2.8 for reading – roughly equal to 4 percent) can hence be attributed to a selection effect that the first analysis was not able to control for.

[11] Calculation: $-0.966[\text{coefficient}] \times 2.38[\text{mean absolute deviation in \%}]/1.14[\text{sd of test score}]$.

Table 6.4: Main results – impact of the educational expansion on students' test scores

	Competence scores							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\ln(\#teachers_{stF}) \times D$	-0.822^{**}	-0.822^{**}	-0.816^{**}	-0.966^{**}	-0.966^{**}	-0.968^{**}	-0.969^{**}	-0.966^{**}
	(0.382)	(0.382)	(0.0383)	(0.378)	(0.378)	(0.381)	(0.381)	(0.381)
$\ln(\#teachers_{stF})$	0.332	0.341	0.321	0.229	0.289	0.033	0.058	0.335
	(0.413)	(0.417)	(0.410)	(0.358)	(0.347)	(0.310)	(0.307)	(0.399)
Subject FE		✓	✓	✓	✓	✓	✓	✓
Gender			✓	✓	✓	✓	✓	✓
School state FE				✓	✓	✓	✓	✓
Cohort FE					✓	✓	✓	✓
Wave & class FE						✓	✓	✓
Test month FE							✓	✓
State specific trends								✓
Observations	10,330							
Number of pupils	6,772							
Number of teachers	322							

Federal-state-by-year-level clustered standard errors in parentheses, * $p < .1$, ** $p < .05$, *** $p < .01$. Baseline regression equation is shown in (6.2). All columns refer to a separate regression with additional Federal State and year fixed effects plus all effects indicated.

How do the effects place themselves in the literature? Chetty et al. (2014a) use an event study of teachers who move between schools as a natural experiment to assess the impact on the test scores of the newly taught students. They find that test scores are raised by 3.5 percent of a sd because of the entry of a teacher from the top 5 percent of the teacher value-added distribution (as assessed by data on previous years). On the one hand, Eq. (6.3) shows that the effects in β_{FE} and β_{DiD} are already adjusted to the educational expansion teachers (by p_2/p_1, see Eq. (6.3)). On the other hand, it is not adjusted to the average quality differential between marginal and non-marginal EETs. Because this differential is most likely to be lower than between a teacher from the top 5 percent versus an average teacher, the β_{FE} and β_{DiD} needs to be inflated. This fact puts my results even more into the range of the findings of Chetty et al. (2014a). The results presented here are in that sense reduced-form effects, since I am not able to normalize them using value-added measures (as the second stage of a two-stage-least-squares estimation would do).

If I expect the same effects as in Chetty et al. (2014a) to operate in my data (0.14sd for math and 0.10 for English:[12] 0.12 on average), I can back out a first stage: the effect of an expanding teacher force on teacher quality (the quality differential in Eq. (6.3)). In

[12]The scores are normalized on a one-sd increase in the teacher value-added.

this case, every 1 percent increase in p_2/p_1 would induce individuals to become teachers such that the value-added of the whole teachers' cohort is increased by 0.0805sd.[13] Also, the literature offers estimates on such a "first stage." Nagler et al. (2015) aim at estimating the effect of recessions on teacher quality, which may be roughly comparable to this setting. They find that due to a recession, the teacher value-added increases by 0.11sd in math and 0.05sd in reading for recession teachers. On average, this is equivalent to the back-of-the-envelope calculation that also yields 0.08.

6.5.2 Assessing the validity of the estimates

Threats to the identifying assumptions

To check that the overall effects are not driven by anything but the causal effect of the subject teacher on the subject test score, I present two complementary pieces of evidence in Table 6.5.

Table 6.5: Robustness checks – placebo regression and predicting parental characteristics

	Placebo regression		Parental characteristics		
	Math teacher – reading score	German teacher – math score	log HH income	Edu. years	
				mother	father
	(1)	(2)	(3)	(4)	(5)
$\ln(\#\text{teachers}_{st})$	0.079	0.268	−0.085	−0.236	−0.110
	(0.514)	(0.516)	(0.399)	(0.353)	(0.360)
Observations	2,713	2,437	2,361	4,079	2,749
Number of teachers	168	158	226	343	315

Federal-state-by-year-level clustered standard errors in parentheses, $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$. All columns refer to a separate regression with federal state and year fixed effects.

First, I present a placebo regression where I assign to each teacher the cross-subject test score; hence, reading scores to math teachers and vice versa (put differently, regression model (6.1) is estimated within each of the light gray cells in Table 6.1). Results of this placebo regression are presented in the first two columns of Table 6.5. If at all, having a math teacher who took up the profession because of the educational expansion raises the reading competence scores of his students (column 1).

[13] The exact calculation looks like this:

$$\underbrace{\text{Second Stage}}_{\text{from Chetty et al. (2014a)}} = \frac{\overbrace{\text{Reduced Form}}^{\text{from Table 6.4}}}{\text{First Stage}}$$

$$\Leftrightarrow \text{First Stage} = \frac{0.00966 \left[\frac{\text{Test score}}{1\,\%\text{ increase in \# teacher}_{st}}\right]}{0.12 \left[\frac{\text{Test score}}{\text{Teacher value-added}}\right]} = 0.0805 \left[\frac{\text{Teacher value-added}}{1\,\%\text{ increase in \# teacher}_{st}}\right]$$

Similarly, this kind of teacher in German does not decrease his student's math competence score (column 2). This finding is consistent with the notion that teachers affect the test score mainly in the subject they teach. Thus, there is not much evidence of either a school selection effect or a spillover effect.

Second, an implicit assumption of the regression models (6.1) and (6.2) is that – conditional on all controls, foremost the fixed effects – everything apart from the educational expansion rate of the teacher is held fixed, even potential factors that are not incorporated in the regression (see Pei et al., 2017 for details). To test for this, I consider potentially important predictors for students' learning outcomes: their socioeconomic background measured by the log household income of the parents as well as the years of education of both the fathers and mothers. If, in a pooled regression (math and German teachers), the teachers' educational expansion rate at the time of his high school graduation is able to predict the parental background of the teachers' students, at least part of the effect could be put into question. In this case, it would not be sufficient to control for the parental background, as further important variables that are still left out of the regression are easily conceivable. Results of this analysis are presented in the last three columns of Table 6.5. It shows that changes in $\ln(\#\text{teacher}_{st})$ have neither the power to predict the household income of the student (column 1), nor years of education of the mothers (column 2) or fathers (column 3). Hence, both supplementary analyses support a causal interpretation of the effects of β_{FE} presented in Table 6.3. It should be noted, however, that math teachers have a marginal impact on reading competence – even more so vice versa. Additionally, EET also teach pupils from a marginally more adverse background.

A caveat may be teacher non-response if it is correlated with the educational expansion rate. Table 6.13 shows that teachers who are willing to provide some background information also teach students that score higher in the math and reading tests. However, this effect disappears once it is conditioned on school fixed effects. This finding suggests that school principals and peer pressure may mainly enforce participation. Using the main specification (6.2), the consent of the subject teachers is not at all able to predict the scores in his subject. Thus, teacher non-response is an argument to prefer the difference-in-differences over the fixed effects model.

A further concern – that may apply to the fixed effects as well as to the difference-in-differences setup – might be the sensitivity of the effects with regard to the assignment year. Figure 6.10 evaluates the sensitivity of the effect with regard to changes in the assignment year. As it reveals, the conclusion and interpretation of the results does not depend on the exact assignment year. The effects are stable over the range where individuals usually make their job decision. Outside of this range (for instance, before age 15 and after age 25) effects disappear. Lastly, the results are insensitive to the size of the class that the teacher teaches (Table 6.14) and the class size and fraction of students with valid test scores are uncorrelated with the educational expansion rate of the teacher (Table 6.15).

The expansion in tertiary education and its relation to the quality of teacher training

As Kamhöfer et al. (2017) demonstrate, the educational expansion also massively affected the university landscape of Germany (from 1962 to 1990, the number of universities doubled). Hence, it is legitimate to ask whether the potential teacher quality differential underlying the main results stems from a difference in the quality of the teacher training in newly opened universities. Table 6.6 therefore presents evidence on whether quality differentials at the university level are a relevant driving force. To check whether factors on the university side are driving the results, I rerun the most saturated specification from Table 6.4 (presented again in column 1 of Table 6.6) and further add university fixed effects (column 2).

Table 6.6: Driving force behind effect

	β_{DiD}		
	(1)	(2)	(3)
$\ln(\#\text{teachers}_{stF}) \times D$	-0.966^{***}	-0.727^{*}	-1.231^{***}
	(0.381)	(0.368)	(0.381)
$\ln(\#\text{teachers}_{stF})$	0.335	0.319	0.407
	(0.399)	(0.435)	(0.408)
Teachers' university fixed effects		✓	
Teachers from new universities dropped			✓
Observations	10,330		9,156
Number of teachers	322		281

Federal-state-by-year-level clustered standard errors in parentheses, $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$.

As shown in column (8) of Table 6.4

Although the magnitude of the effect shrinks by about one quarter in absolute terms, the effect remains significant and economically relevant even after absorbing a potentially high fraction of the identifying variation. Thus, the result indicates that heterogeneity in university quality only explains a small fraction of the effect. But openings can also lead to a selection of high-ability individuals becoming teachers. To check this, I drop teachers that graduate from new universities and re-estimate Eq. (6.2). The resulting estimate is higher and thereby provides some evidence that university openings generally induced teachers of a higher quality to enroll in teacher training.

6.5.3 Detecting teacher selection in the characteristics of teachers

So far, I looked at whether teachers have a different ability (i.e., teacher quality) to raise the test scores of their students with respect to different degrees of their exposure to the educational expansion. Although this is considered to be the ultimate measure of teacher quality (see, e.g., Hanushek and Rivkin, 2006 or Chetty et al., 2014b),

one can still ask whether the teachers not only have a better quality but also different characteristics that are correlated with quality (Jackson et al., 2014). This serves two purposes. First, if I found effects, this would strengthen the credibility of the main effects on test scores. And second, it is important for tailoring future policies, since hiring decisions or enrollment conditions for prospective teachers may be based on characteristics that correlate with teacher quality. The NEPS data set provides additional information on teachers. In addition to the birth year and the federal state of high school graduation that was used throughout the analysis, the data also includes the grade of high school and university graduation. In addition, the data contains subjective indicators that are targeted to retrospectively portray aspects of the reasons why they became teachers. Ten questions in the questionnaire for teachers try to capture these aspects. Teachers have to assess the relative, subjective importance of each aspect on a four-point Likert scale (ranging from very unimportant, 1, to very important, 5). For two reasons, it may be suboptimal to present estimates on all 10 domains. First, multiple testing may be a concerning issue, since one cannot determine at which domain to expect an effect and on which not a priori. Second, teachers may differ generally in their answer patterns. For instance, low-quality-teachers may place a higher importance on all domains in general. High-quality teachers may tend to place less weight on all domains but relatively more on those that correlate with intrinsic motivation. Those two opposing patterns may then confound the overall effect.

I therefore conduct a factor analysis that serves to detect these patterns. This is similar to Rockoff et al. (2011) who employ variables on cognitive skills. Because I expect two latent factors to be inherent in the answer patterns – namely intrinsic and extrinsic motivation – I opt for a principal component analysis with two factors.[14] For the 10 questions, the resulting two factor loadings are plotted in Figure 6.6. The horizontal axis maps the first dimension and the vertical axis the second factor loading. The loadings on the first domain are all positive. This can be ascribed to a general positive correlation between all of these subjective questions. This general correlation is purged out of the second loading. Therefore, it may be more informative for the analysis. Indeed, the second domain clearly shows that the variables form two clusters. Specifically, the importance of leisure, salary, job security, the prestige of the job, and being able to reconcile the job with a family life form one cluster (positive factor loading). Since all those domains are not specific to the teacher profession, I refer to these variables as those reflecting external motivation. The remaining variables have a negative factor loading. These variables comprise the joy to teach, the challenges of the job, being around people, the dedication to the subject and to accomplish certain goals in the job. The common feature of these variables is that they are all job-related. Hence, this cluster reflects intrinsic motivation. These two clusters are present in the latent correlation of the variables. It is crucial whether the scores formed by those factor loadings are affected by the exposure to the educational expansion ($p2/p1$). If

[14] Principal component analysis simply transforms p-dimensional data into $m < p$ dimensional data, where p is the number of principal components along which the data varies most. Technically, the first component is a summary score of the data $PC_1 = \phi_{11}x_1 + \phi_{21}x_2 + \cdots + \phi_{101}x_{10}$ and ϕ_{i1} are the factor loadings of the first component. The ϕ's are chosen such that they maximize the sample variance of PC_1 under the constraint that $\sum_{i=1}^{10} \phi_1 i^2 = 1$. The second principal component PC_2 again maximize the variance of the data, but with the additional condition that PC_2 is orthogonal to PC_1.

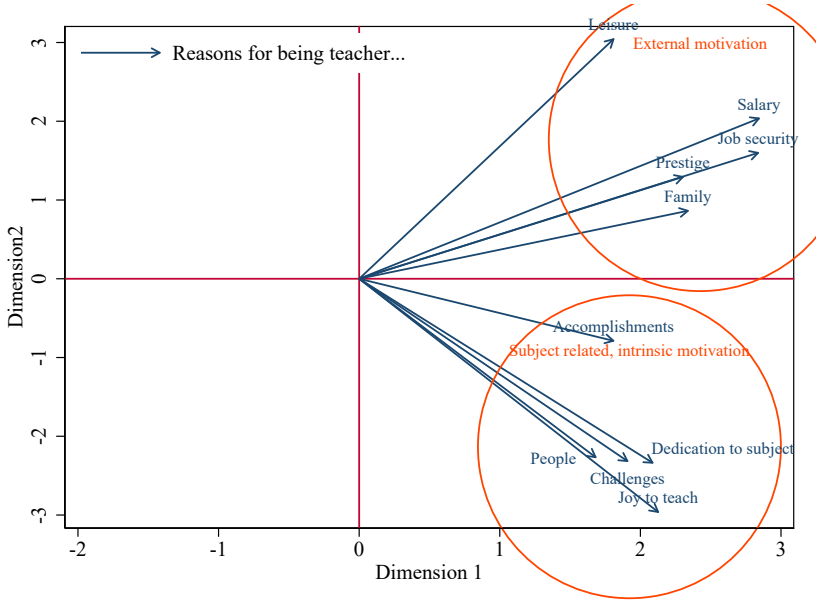


Figure 6.6: Cluster analysis of aspects teachers' job choice

Notes: the graph (biplot) plots the factor loadings resulting from a principal component analysis with two components on 10 variables that capture the aspects of the job choice of the teachers.

the scores and $p2/p1$ were correlated, this would indicate that EETs have a different kind of motivation.

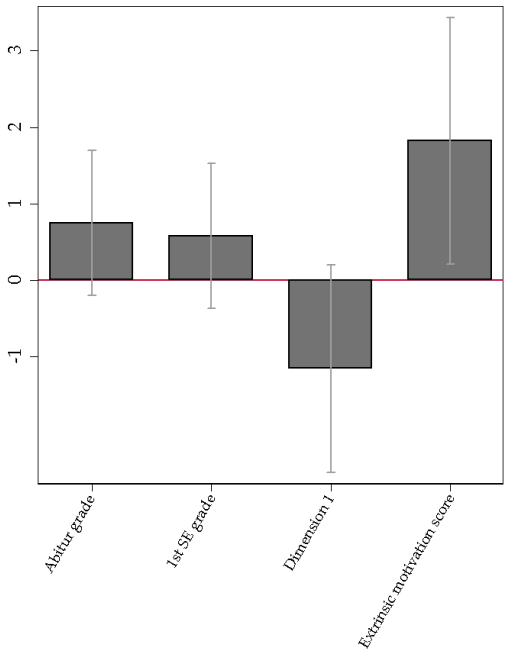


Figure 6.7: The educational expansion and teachers' characteristics

Note each bar depicts the effect of $\ln(\#\text{teachers}_{st})$ on the outcome indicated by the label below the bar. The sample is equivalent to the difference-in-differences regression with 322 teachers, standard errors are adjusted on the teacher level. Confidence bands indicate the 90% confidence level.

The bar plot in Figure 6.7 presents evidence on this. Each bar represents the effect of $\ln(\#\text{teacher}_{st})$ on a respective outcome variable (indicated by the label below each bar). The sample is equivalent to the DiD regression. As before, standard errors remain adjusted to the federal state-year level. The first bar shows that the higher p_2/p_1, the worse the grade of academic track high school graduation (Abitur). Hence, this effect indicates that teachers with worse high school grades take up the teaching profession at times of high demand for teachers. The effect, however, fails to be significant at the 10 percent level. Does this hint at the lack of statistical precision or point to negligible economic meaning? Table 6.12 tries to shed light on this question by comparing the coefficients of a teacher-level regression of different samples. It turns out that the coefficients are stable, irrespective of whether academic track teachers from other subjects without assignable student test scores are included (column 2) or middle school (Realschule) teachers are further added (first column 1). But statistical precision increases by adding more teachers. These results for the German educational expansion are in contrast to findings for the US where no powerful predictors of teacher quality are identified (Jackson et al., 2014).

Returning to Figure 6.7, this figure further shows that the effect on the high school grade also propagates to university. Here, EETs have marginally worse grades. Beyond grades, is there evidence that teachers affected by the education expansion have a different work ethic? The third and fourth bars shed light on this by analyzing the principal component summary measures. The former shows that EETs tend to generally place significantly less importance on all domains captured by the questions because the first domain places almost equal and positive weight on all the domains. One explanation for this effect is a potentially different reference point of those teachers. Yet, distinguishing these questions as suggested by the second dimension is more informative. On this dimension, I find that there is a positive effect for EETs. This means that EETs place significantly more weight on questions with a positive weight (the external motivation to be a teacher) and less on those with a negative weight (the intrinsic cluster of the questions). This finding suggests that EETs have a slightly shifted work ethic from intrinsic to extrinsic motivation, which is compatible with the the main-effect: EETs may not put as much effort into raising the test scores of the students because they do not gain their motivation from it.

6.6 Conclusion

This paper shows that more teachers do not mechanically lead to smarter students. It thereby emphasizes an important mechanism of quickly expanding public institutions: focusing on quantitative aspects may deter quality, all else equal. This message can also be important for today's objective to increase the scale of public institutions, such as the current expansion of daycare facilities for children (BMFSFJ, 2015).

Using one of the major social changes in the past 60 years, the educational expansion, I test whether this social change attracted individuals with a different quality to eventually become a teacher. With the help of a simple expression of how the group average of teacher quality changes in response to newly entering teachers, the effect can be placed into the literature of broadly related studies. In a baseline fixed effects setup

the impact of the educational expansion on teachers is separated from teachers' experience and federal state-specific effects. To take care that no school selection effects impact the results, I estimate a between-subjects difference-in-differences model.

The evidence I get from this approach suggests that the average effect of the educational expansion, which caused teacher quality to diminish, was roughly 2 percent of a standard deviation in students' test scores (math and reading). Comparing this ("reduced-form") effect to existing studies on teacher selection (e.g., Nagler et al., 2015 who provide a surrogate of a "first stage") and with the effect of teacher quality on students' test scores (Chetty et al., 2014a, a "second stage") the results of this paper are well-placed into the existing literature on the US. Thus, the scope of the effects on the students in this study are likely to also extend to labor market performance in adulthood. These results are further substantiated by the findings that teachers who are selected because of the educational expansion performed better at high school (though not at university) and have a slightly different work ethic that is based more on extrinsic rather than intrinsic motivation.

Potential policy implications are non-trivial, since not expanding the higher secondary education would not have been a solution either. Nonetheless, policy could very well have reduced its demand for teachers while sticking to the provision of a sufficient amount of spots for students. Taking the evidence of this paper together with a further characteristic of the educational expansion – the student-teacher ratio that declined at the same time in Germany (depicted in Figure 6.8a) – the educational policy departments of the federal states may have attenuated this tradeoff between teachers and students' learning outcomes by not simultaneously pushing down the student-teacher ratio. Either granting more pupils access to higher secondary education while increasing the student-teacher ratio at the same time or focusing more on investing in quality. For instance, this can be done by improved teacher training or via a more selective process of hiring teachers.

6.7 Appendix

Assumptions of the difference-in-differences model

The underlying assumptions of this approach are threefold. The first assumption is that teacher quality matters similarly for math courses as it does for German courses: $Y^{\text{Math}}\left(Q^{\text{Math}}\right) \approx Y^{\text{Reading}}\left(Q^{\text{German}}\right)$, where $Y^u(Q^v)$ refers to the potential test score of a student in subject u which might depend on the latent teacher quality Q of a teacher who teaches the student in subject v. This assumption is important for the interpretation of the effect.[15] In the same vein, the second assumption rules out which effect I do not expect to see. If $u = v$ (a teacher in a certain subject can only affect

[15] If there is a structural difference between the subject-specific effects of teacher quality the identified effect of (6.2) would be a weighted average which would change the economic interpretation of β_{DiD}.

the test scores of her students in the same subject) I expect to see an effect else it can be ruled out.

$$Y^{\text{Reading}}\left(\mathrm{Q}^{\text{Math}}\right) = Y^{\text{Reading}} \quad \wedge \quad Y^{\text{Math}}\left(\mathrm{Q}^{\text{German}}\right) = Y^{\text{Math}}$$

Those two assumptions allow me to precisely define a treatment indicator that indicates whether the teacher's subject ($v_j \in \{1,2\}$) is the same as the test score under consideration ($u_i \in \{1,2\}$).

$$\begin{aligned} D &= \begin{cases} 1 \quad \text{if} \quad (\overbrace{\text{test=Math}}^{u_i=1}) \quad \wedge \quad (\overbrace{\text{teacher=Math}}^{v_j=1}) \\ \qquad \vee \; (\underbrace{\text{test=Reading}}_{u_i=2}) \quad \wedge \quad (\underbrace{\text{teacher=German}}_{v_j=2}) \\ 0 \quad \text{else.} \end{cases} \\ &= \mathbb{1}\left(u_i = v_j\right) \end{aligned}$$

Also, these assumptions enable us to redefine the potential outcomes as Y^1, Y^0 in order to reconcile it with the treatment indicator.

The third assumption is actually most crucial for identification, since it states which variation in the response variable can be causally attributed to variation in the gradual changes in the measure of the educational expansion. To be more precise, I assume that the quality differential of any pair of math and German teachers in the same class is independent of the potential test scores of their pupils:

$$\left(Y^1(\mathrm{Q}^1) - Y^1(\mathrm{Q}^0)\right) \perp\!\!\!\perp \left(\mathrm{Q}^1 - \mathrm{Q}^0\right) \mid \mathbf{X}_{\text{FE}} \tag{6.4}$$

where $\mathbf{X}_{\text{FE}}$ comprise teacher year and federal state fixed effects and class fixed effects. This assumption may be credible, as parental background, class, and school effects, and any further individual differences are held fixed. It would be violated, e.g., if the within-class variation in potential test scores is large, which school principals could observe alongside the quality of their teachers. In addition they had to strategically assign teachers (and their quality) to courses and classes such that test scores between courses are, for instance, either compensated or reinforced between subjects. In this case, at least some parts of β_{DiD} in (6.2) would also capture a selection effect. This, however, is unlikely to dominate the effect, since within classes it appears more plausible that relative advantages in one particular subject cancel each other out. Although I term this strategy differences-in-differences, the argumentation above clarifies the analogy to an instrumental variables approach, where the school principal's assignment is the plausible random assignment mechanism that I exploit for identification.

Robustness of the the employed educational expansion rate

As federal state and year fixed effects are used as (the most important) control variables, the main coefficients of interest, β_{FE} and β_{DiD}, are essentially identified by changes from year and federal state-specific means: $d\ln(\#\text{teachers}_{st})$. Instead of using $\ln(\#\text{teachers}_{st})$ as the regressor of interest, one could equivalently have used the residual from the following regression (Frisch-Waugh-Lovell Theorem):

$$\ln(\#teachers_{st}) = \delta_s + \gamma_t + u_{st}$$

, this residual u_{st} equals $d\ln(\#\text{teachers}_{st}) = {}^{d\#\text{teachers}_{st}}/_{\#\text{teachers}_{st}}$, which essentially is the ratio of educational expansion teachers to the projected number of teachers needed in the absence of the educational expansion. Using the notation from Section 6.3.2, this is ${}^{p_2}/_{p_1}$. But as we have seen from Eq. (6.3), ${}^{p2}/_{(p_1+p_2)}$ is considered to be the appropriate leverage by which the average teacher quality of a certain teacher cohort in a federal state is affected by the average quality of the incoming teachers. Thus, does ${}^{p2}/_{(p_1+p_2)}$ better capture the quality effects? To check this, one can adjust the residual u_{st} (relative change in the teacher force with respect to the projected number of teachers) to the relative change with respect to all teachers by dividing by $\theta = 1 + {}^{p_2}/_{p_1}$. [16] Plugging in ${}^{u_{st}}/_{\theta}$ instead of u_{st} in regressions 6.1 and 6.2 yields estimates that are presented in Table 6.7.

The results presented in this Table indicate that all specifications are largely insensitive toward whether ${}^{p2}/_{p_1}$ or ${}^{p2}/_{(p_1+p_2)}$ are employed in the regressions. Hence, β_{FE} and β_{FE} indeed seem to adjust the effect to the marginal teachers.

Tables

[16] The parameter θ can be derived as follows:

$$\begin{aligned} \theta\frac{p_2}{p_1+p_2} &= \frac{p_2}{p_1} \quad \Leftrightarrow \quad \theta p_2 p_1 = p_2(p_1+p_2) \\ \Leftrightarrow \quad \theta &= 1+\frac{p_2}{p_1} \end{aligned}$$

Table 6.7: Tansformed results

	β_{FE}		β_{DiD}
	Math	Reading	Pooled
	(1)	(2)	(3)
Adjusted measure: u_{st}/θ	-1.372^{***}	-0.708	0.953^{***}
	(0.520)	(0.481)	(0.380)

Notes: Federal-state-by-year-level clustered standard errors in parentheses, $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$. This table assesses whether the main effects in Tables 6.3 and 6.4 are adjusted appropriately to induced changes on the average "quality" of teachers by incoming educational expansion teachers. The identifying variation plotted in Figure 6.3 is p_2/p_1, but the leverage of educational expansion teachers on the average teacher quality of a cohort of teachers from federal state s in year t is p_2/p_1+p_2, as shown in Eq. (6.3) . Therefore, the identifying variation (the residual from a first-stage regression) is divided by the factor $\theta = 1 + p_2/p_1$ and plugged into a second-stage regression.

Table 6.8: Teacher mobility between federal states

	Number of teachers	Percentage
Teacher does not move	234	73.9
Teacher moves to neighboring states	38	11.8
Teacher moves to non-neighboring states	46	14.3
Total	318	100

Notes: teacher mobility is defined as whether a teacher is employed at a school in a federal state that is different to the federal state in which the teacher graduated from high school.

Table 6.9: Example structure of the data for the triple differences estimation

Student level				Treatment	Teacher level					
ID	Name	Subject	Test score		ID	Name	Subject	Year teacher turned 19	Federal state	$d\ln(\#\text{teacher}_{st})$
1	Alexander Meier	Math	2.34	1	1	Lothar Müller	Math	1980	Bavaria	0.08
1	Alexander Meier	German	1.65	0	1	Lothar Müller	Math	1980	Bavaria	0.08
1	Alexander Meier	German	1.65	1	2	Esther Schulz	German	1978	Hesse	-0.02
1	Alexander Meier	Math	2.34	0	2	Esther Schulz	German	1978	Hesse	-0.02

Notes: This table shows the structure used to estimate the main results of the paper by a triple difference estimation. Her pupil is observed four times – two observations for each subject-specific test score (math and reading) by math (here Lothar Müller, lines 1-2) and German teacher (Esther Schulz, lines 3-4). Within each teacher, the test score outcome of each assigned pupil that relates to the respective subject of the teacher serves as a treatment whereas the other test score serves as the control group. To difference out any subject to account for subject-specific effects, the data are expanded on the pupil level such that treatment and control group are reversed. This expansion of the observations, the standard errors remain clustered at the year the teacher turned 19 and the federal state of the teachers' high school graduation.

Table 6.10: Number of pupil and number of students used in this analysis and dropping reasons

	(1) Plain sample	2) Higher secondary track students	(3) Western Germany	(4) Gymnasium track	(5) Math or German teachers	(6) Teacher older than 19 in 1970	(7) Math teachers	(8) German teachers
N	87,776	53,745	17,337	10,806	5,683	5,311	2,625	2,855
# Students	24,417	14,532	9,042	5,345	4,259	3,980	2,491	2,620
# Teacher	4,952	2,779	1,206	745	345	322	158	168

Notes:
Students of either Realschule, Gesamtschule or Gymnasium.

Table 6.11: Descriptives for aspects of teacher's job choice

	Statistics	
	Mean	SD
Reconcilability of job and family	3.353	(0.778)
Possibility to interact with people	3.573	(0.538)
Leisure time	2.125	(0.786)
Salary	2.630	(0.737)
Meet challenges	3.013	(0.669)
Joy to teach	3.691	(0.489)
Job security	3.200	(0.750)
Prestige of being teacher	1.863	(0.776)
Possibility to accomplish things	2.382	(0.759)
Dedication to subject	3.580	(0.553)

Domains of job choice are based on answers on the following question: "How important was the following aspect for your choice of becoming a teacher?" Teachers could respond on a 5-point Likert scale ranging from 1 – "Very unimportant" – to 5 –"Very important".

Table 6.12: The association between the degree of the relative degree of the educational expansion and the Abitur grade for different samples

	Grade Abitur		
	(1)	(2)	(3)
$\ln(\#\text{teacher}_{st})$	0.942^{***}	0.970^{**}	0.956
	(0.362)	(0.419)	(0.661)
Sample restrictions:			
-Realschule	✓		
-Gymnasium	✓	✓	
-Sample teachers	✓	✓	✓
# teacher	995	625	284

Standard errors in parentheses, $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$. Each column shows the effect of the educational expansion on the selection of teachers indicated by their grades. The underlying data is on the teacher level. Control variables comprise year fixed effects, federal state fixed effects and subject fixed effects.

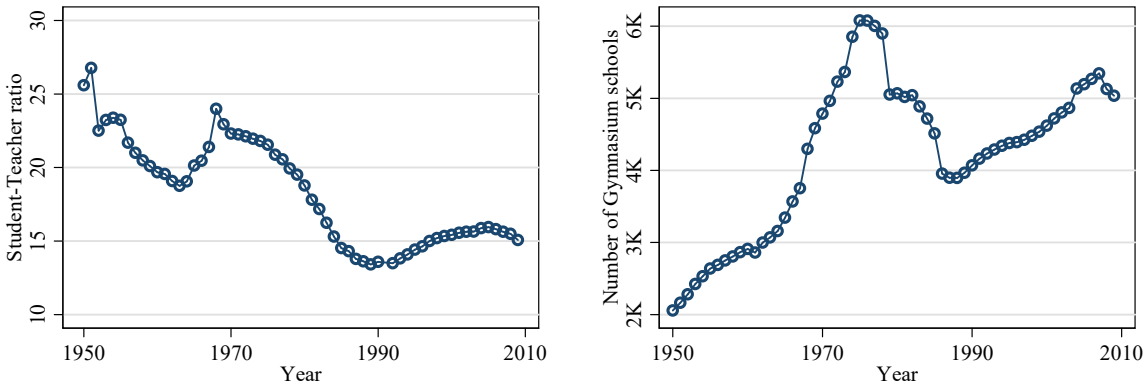
Table 6.13: Potential impact of teacher non-response on the main effect

	Test scores							
	Math			Reading			Pooled scores	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Valid Teacher	0.138***	0.087***	0.040	0.136***	0.079***	0.032	0.136***	0.137***
	(0.035)	(0.030)	(0.030)	(0.032)	(0.026)	(0.028)	(0.032)	(0.027)
Valid Teacher × D								0.0004
								(0.027)
Cross-subject score		✓			✓			implicitly
School fixed effects			✓			✓		
# observations	15,123	14,621	15,123	14,831	14,621	14,831	29,954	29,954

Notes: This table shows whether information on teacher having a valid interview (a prerequisite for knowing his birth year and the federal state of high school graduation among others) is able to predict the test scores of his students. Columns (1) to (6) report regression coefficients from the fixed effects regression similar to (6.1) whereas columns (7) and (8) show the results of the difference-in-differences regression as reported in (6.2). Most importantly, this table shows that while there may be some bias from teacher non-response in the fixed effects strategy (6.1) (it is still unclear whether this is correlated with the educational expansion), this bias disappears once within-school variation is used (columns (3) and (6)). If the difference-in-differences strategy is employed meaning that within-course variation is solely and effectively exploited, teacher non-response is not at all able to predict the test scores of the students.

The reading score for math teachers and math test score outcomes and math scores for German teachers and reading score outcomes.

Figures



(a) Student-Teacher ratio over time (b) Number of schools over time

Figure 6.8: Further characteristics of the educational expansion in Germany

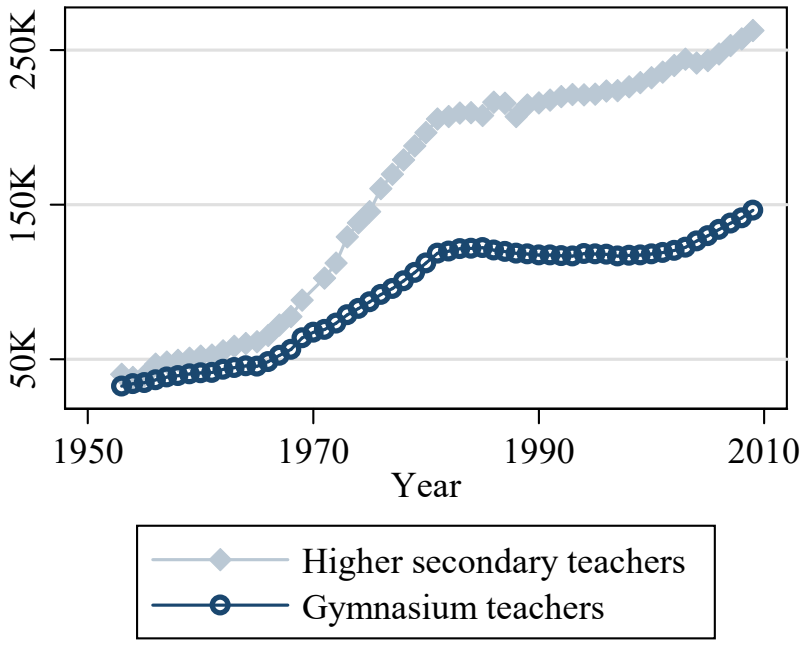


Figure 6.9: Number of Gymnasium teacher over time

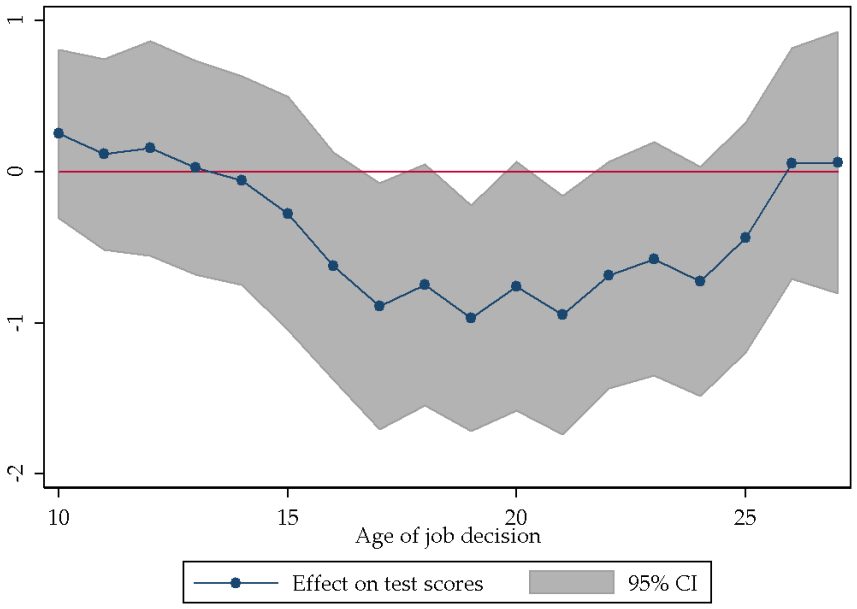


Figure 6.10: Sensitivity of the effect with respect to the assignment year

This graph plots the effect of EET on student test scores (β_{DiD}) and how this effect changes with respect to a different assignment year. In the main analysis, the assignment year was set to 19, that is, the year of academic track education. The effects are similar in magnitude and precision over the age range from 17 to 21. Outside this range, the effect is negligible.

Supplementary material

Table 6.14: Sensitivity of the results due to selective student non-response and test participation

	(1)	(2)	(3)
	Baseline	Class size FE	Class size≥10
$\ln(\#\text{teachers}_{st}) \times D$	-0.966^{**}	-0.961^{**}	-1.0611^{**}
	(0.381)	(0.383)	(0.415)
$\ln(\#\text{teachers}_{st})$	0.335	0.231	0.344
	(0.399)	(0.409)	(0.498)
D	8.384^{**}	8.341^{**}	9.201^{**}
	(3.292)	(3.305)	(3.582)
# observations	10,330	10,330	9,465
# teachers	322	322	251

Notes: The coefficients are estimated using equation (6.2). Column (1) refers to the results from column (8) in Table 6.4. Column (2) includes class size fixed effects to see whether the main component of the correlation structure has to be attributed to the class size that may correlate with $\ln(\#\text{teachers}_{st})$. Here, "Class size" refers to the number of students with a valid test score observation. To further see whether the main result is actually driven by very small classes, column (3) drops classes with less than 10 students.

Table 6.15: Potential impact of student non-response on the main effect

	Test score class size	Fraction with valid test score
	(1)	(2)
$\ln(\#\text{teachers}_{st}) \times D$	-0.877	0.014
	(1.761)	(0.112)
$\ln(\#\text{teachers}_{st})$	-13.410^{***}	-0.673^{***}
	(5.955)	(5.369)
# observations	10,330	6,393
# teachers	322	208

Notes: The coefficients are estimated using equation (6.2). Test score class size refers to the number of valid test score observations by course (math or German). Fraction with valid test score observations divides the the test score class size by the actual number of students per class. Its observation number is lower because of the non-response of the respective class teacher.

Part IV

Concluding remarks

Conclusion

This dissertation is about the individual consequences of two of the most recent changes to industrialized societies – the demographic change and the educational expansion. Both can be characterized and quantified by statistics that exhibit substantial trends. The former, for instance, can be measured by population aging, while the latter can be described by increasing enrollment or graduation rates in higher education, such as college education.

We have seen that these transformations affect individual decisions that entail important individual consequences, which, in turn, may constitute public challenges. The first chapter demonstrates that rather mild variations in the intra-uterine disease environment may lead to effects that not only affect the life of the fetus but also transmit to the next generation as hypothesized by the fetal origins hypothesis. From a policy point of view, the results show the importance of investments in maternal health. The returns on investments in prenatal maternal care may have very long-run returns and thereby may affect societal inequalities even in the next generations.

The findings of the subsequent two chapters of this dissertation indicate that informal care goes along with higher societal costs as commonly expected. Specifically, I showed that there are considerable short-term effects on mental health that, however, fade out over time while physical health seems to be unaffected by caring informally. These results appear to be independent of the length of the care episode. Beside full-time employment, the results also show significant wage effects that seem to form rather in the long-run. In total, the labor supply effects have more immediate fiscal consequences as compared to the health effects. A simple back-of-the-envelop analysis estimates these fiscal costs to roughly amount to € 860 million annually. If taken at face value, these costs together with a monetary assessment of the health effects should be taken into account when debating over expansions in either the formal or informal care sector. In total, this dissertation may therefore provide the policymaker with some valuable ideas on how to shape future policies.

The results of Chapter 4 and 5 show that college education has sizeable individual consequences that may also cause changes to the society from a macro perspective. On average, the results show positive nonmonetary returns to college on physical (but not mental) health and cognitive skills. These effects seem to be driven by more demanding jobs that slow down the cognitive decline and better health behaviors. The structural heterogeneity underlying these results reveals that those with the highest preferences for college education also have the highest returns. Therefore, while more education can prevent individuals from suffering from dementia (assuming the cognitive reserve hypothesis to be correct), the policy implications are less encouraging since only 30–40 percent of the population exhibits these potential positive college

returns. A future extension of the college landscape is likely to encourage only those individuals to go to college that have much smaller returns.

Also when assessing female fertility decisions, we find heterogeneous results. The results show that there are basically two groups of women, those who seem to prefer a career over a family and those who have reverse preferences. The findings are consistent with having these two groups of women: there is a negative effect on the intensive, but even a slightly positive effect on the intensive fertility margin. In addition, the college returns for women versus mothers also reveal a heterogeneity that is in line with these preferences. If the policy maker would like to increase the fertility of highly-educated women, the implications are clear: most appropriate appears the often-debated compatibility between work and family affairs, such that also career-type women can have children without lowering their sights.

Lastly, Chapter 6 showed that policy-induced social changes might incur side-effects. In the analyzed setting, I evaluated the policy-induced expansion of the teacher force in the higher secondary education in Germany (Gymnasium). This expansion might have had consequences that are still detectable in the test scores of today's students'. Therefore, expanding the public sector may have adverse long term consequences. However, also some observed characteristics are associated with these educational expansion teachers are different and observable already at the time of the expansion. This finding emphasizes a potential leverage of the policymaker. In my setting this was, for instance, the highschool exit grade of prospective teachers. If the hiring decision of the teachers were more strictly based on this characteristic, potentially, these side-effects would have been avoided.

In a nutshell, knowledge about these individual consequences that are caused (directly or indirectly) by social changes is pivotal for shaping future policies that are supposed to alleviate many of the imposed challenges. These policies may include finding and promoting the efficient and economic mode of long-term care provision on the supply side. On the demand side, preventing individuals from becoming care dependent and facilitate people to stay healthy longer is a key factor. This might be done by, for instance, promoting education. Further leverages of public policy to mitigate the effects of the population aging may comprise raising the low fertility rates of high-educated women. Yet, implementing the policies above lopsidedly or too quickly may incur potential side effects. One example may be the expansion of secondary education, which was partly policy-induced. Insights into the prevalence of diminishing quality of this public institution while increasing its scale and scope is likewise precious for the policymaker. From a more general perspective, knowing how persistent societal inequalities are modified due to social change in the long run can also be important for complementing the evidence on how societies react to social change.

Summing up the general results of this dissertation, it highlights the general need for action of the policymaker that is associated with the analyzed past changes and future changes. This dissertation points out causal evidence that address some of these challenges and thereby may have important policy implications. Although this thesis may contain some contributions to the literature, many are left for future work. These are related in particular to the understanding of the driving forces behind the

identified effects. Understanding exactly how education operates on health and cognitive skills or health is mentioned as just one example.

Bibliography

Acemoglu, D. (2002). Technical Change, Inequality, and the Labor Market. Journal of Economic Literature, 40(1), 7–72.

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.

Adda, J., Dustmann, C., and Stevens, K. (2017). The Career Costs of Children. Journal of Political Economy, 125(2), 293–337.

Aizer, A., and Currie, J. (2014). The Intergenerational Transmission of Inequality: Maternal Disadvantage and Health at Birth. Science, 344(6186), 856–861.

Akachi, Y., and Canning, D. (2007). The Height of Women in Sub-Saharan Africa: The Role of Health, Nutrition, and Income in Childhood. Annals of Human Biology, 34(4), 397–410.

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. (2006). Is the 1918 Influenza Pandemic Over? Long–Term Effects of In Utero Influenza Exposure in the Post–1940 U.S. Population. Journal of Political Economy, 114(4), 672–712.

Almond, D., and Currie, J. (2011). Killing Me Softly: The Fetal Origins Hypothesis. Journal of Economic Perspectives, 25, 153–172.

Almond, D., Currie, J., and Herrmann, M. (2012). From Infant to Mother: Early Disease Environment and Future Maternal Health. Labour Economics, 19(4), 475–483.

Almond, D., Edlund, L., and Palme, M. (2009). Chernobyl's Subclinical Legacy: Prenatal Exposure to Radioactive Fallout and School Outcomes in Sweden. Quarterly Journal of Economics, 124(4), 1729–1772.

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

Alzheimer's Association (2017). 2017 Alzheimer's Disease Facts and Figures. Alzheimer's & Dementia, 13(4), 325 – 373.

Alzheimer's Disease International (2013). World Alzheimer Report 2013, Journey of Caring – An analysis of long-term care for dementia. Tech. rep., Alzheimer's Disease International (ADI), London.

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, 127, 171–182.

Anderson, J. (2007). Cognitive Psychology and its Implications. New York: Worth Publishers, 7 ed.

Andreella, C., Karlsson, M., Nilsson, T., and Westphal, M. (2015). The Long Shadows of Past Insults Intergenerational Transmission of Health over 130 Years. Ruhr

Economic Papers 571, RWI Essen.

Angrist, J. D., and Pischke, J.-S. (2008). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Ashraf, N., Bandiera, O., and Lee, S. (2014). Do-gooders and Go-getters: Career Incentives, Selection, and Performance in Public Service Delivery. STICERD - Economic Organisation and Public Policy Discussion Papers Series 54, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE.

Augurzky, B., Hentschker, C., Krolop, S., and Mennicken, R. (2013). Pflegeheim Rating Report 2013 – Ruhiges Fahrwasser erreicht. Hannover: Vincentz network., Essen.

Augurzky, B., Reichert, A., and Schmidt, C. M. (2012). The effect of a bonus program for preventive health behavior on health expenditures. Ruhr Economic Papers 373, Essen.

Autor, D. H. (2014). Skills, education, and the rise of earnings inequality among the "other 99 percent". Science, 344(6186), 843–851.

Bacolod, M. P. (2007). Do Alternative Opportunities Matter? The Role of Female Labor Markets in the Decline of Teacher Quality. The Review of Economics and Statistics, 89(4), 737–751.

Bakx, P., de Meijer, C., Schut, F., and van Doorslaer, E. (2014). Going Formal or Informal, Who Cares? The Influence of Public Long-Term Care Insurance. Health Economics, forthcoming.

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.

Barker, D. J. (1990). The Fetal and Infant Origins of Adult Disease. BMJ, 301(6761), 1111.

Barker, D. J. (1997). Maternal Nutrition, Fetal Nutrition, and Disease in Later Life. Nutrition, 13(9), 807–813.

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 Aapplication to Treatments of Breast Cancer Patients. Health Economics, 16(11), 1133–1157.

Beach, S. R., Schulz, R., Yee, J. L., and Jackson, S. (2000). Negative and Positive Health Effects of Caring for a Disabled Spouse: Longitudinal Findings from the Caregiver Health Effects Study. Psychology and Aging, 15(2), 259–271.

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.

Bejenariu, S., and Mitrut, A. (2013). Austerity Measures and Infant Health. Lessons from an Unexpected Wage Cut Policy. Working Papers in Economics 574, University of Gothenburg, Department of Economics.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. Review of Economic Studies, 81(2), 608–650.

Ben-Porath, Y. (1967). The Production of Human Capital and the Life Cycle of Eearnings. Journal of political economy, 75(4, Part 1), 352–365.

Bengtsson, T., and Dribe, M. (2011). The Late Emergence of Socioeconomic Mortality Differentials: A Micro-level Study of Adult Mortality in Southern Sweden 1815–1968. Explorations in Economic History, 48(3), 389–400.

Bengtsson, T., and Lindström, M. (2003). Airborne Infectious Diseases During Infancy and Mortality in Later Life in Southern Sweden, 1766–1894. International Journal of Epidemiology, 32(2), 286–294.

Bengtsson, T., and van Poppel, F. (2011). Socioeconomic Inequalities in Death from Past to Present: An Introduction. Explorations in Economic History, 48(3), 343–356.

Bhalotra, S., Karlsson, M., and Nilsson, T. (2017). Infant Health and Longevity: Evidence from A Historical Intervention in Sweden. Journal of the European Economic Association, 15(5), 1101–1157.

Bhalotra, S., Karlsson, M., Nilsson, T., et al. (2015). Infant Health and Longevity: Evidence from a Historical Trial in Sweden. Tech. rep., Institute for Social and Economic Research.

Bhalotra, S., and Rawlings, S. (2013). Gradients of the Intergenerational Transmission of Health in Developing Countries. Review of Economics and Statistics, 95(2), 660–672.

Bhalotra, S. R., and Venkataramani, A. (2011). The Captain of the Men of Death and his Shadow: Long-run Impacts of Early Life Pneumonia Exposure. Available at SSRN.

Bharadwaj, P., Løken, K., and Neilson, C. (2011). Early Life Health Interventions and Academic Achievement. American Economic Review.

Bhattacharya, J., and Vogt, W. (2012). Do Instrumental Variables Belong in Propensity Scores? International Journal of Statistics and Economics, 9, 107–127.

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., and Salvanes, K. G. (2007). From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes. The Quarterly Journal of Economics, 122(1), 409–439.

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.

Bleakley, H. (2017). Longevity, Education, and Income: How Large is the Triangle?

Blossfeld, H.-P., Roßbach, H.-G., and von Maurice, J. (2011a). Education as a Lifelong Process – The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft, 14, Special Issue.

Blossfeld, H.-P., Roßbach, H.-G., von Maurice, J., Schneider, T., Kiesl, S. K., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., Prenzel, M. S., et al. (2011b). Education as a Lifelong Process – The German National Educational Panel study (NEPs). Age, 74(73), 72.

BMFSFJ, Bundesministerium für Familie, Senioren, Frauen und Jugend. (2015). Fünfter Breicht zur Evaluation des Kinderförderungsgesetzes – Bericht der Bundesregierung 2015 über den Stand des Ausbaus der Kindertagesbetreuung für Kinder unter drei Jahren für das Berichtsjahr 2014 und Bilanzierung des Ausbaus durch das Kinderförderungsgesetz. Tech. rep.

Boalt, C. (1939). 27.000 måltider: en undersökning av kostvanor. Kooperativa förbundet.

Bobinac, A., van Exel, N. J. A., Rutten, F. F., and Brouwer, W. B. (2010). Caring for and Caring about: Disentangling the Caregiver Effect and the Family Effect. Journal of Health Economics, 29(4), 549–556.

Böhlmark, A., and Lindquist, M. J. (2006). Life-Cycle Variations in the Association between Current and Lifetime Income: Replication and Extension for Sweden. Journal of Labor Economics, 24(4), 879–896.

Bolin, I. (1934). Sockret. Vårt billigaste födoämne. Gothenburg, Sweden, Gø"teborgs litografi AB.

Bolin, K., Lindgren, B., and Lundborg, P. (2008). Your Next of Kin or Your Own Career?: Caring and Working Among the 50+ of Europe. Journal of Health Economics, 27(3), 718–738.

Bölling, R. (1983). Sozialgeschichte der deutschen Lehrer: Ein Überblick von 1800 bis zur Gegenwart. Vandenhoeck & Ruprecht.

Börsch-Supan, A. (2003). Labor market effects of population aging. LABOUR, 17, 5–44.

Börsch-Supan, A., and Jürges, H. (2005). The Survey of Health, Aging, and Retirement in Europe. Methodology.

Bozzoli, C., Deaton, A., and Quintana-Domeque, C. (2009). Adult Height and Childhood Disease. Demography, 46(4), 647–669.

Bozzoli, C., Deaton, A. S., and Quintana-Domeque, C. (2007). Child Mortality, Income and Adult Height. Working Paper 12966, National Bureau of Economic Research.

Brayne, C., Ince, P. G., Keage, H. A., McKeith, I. G., Matthews, F. E., Polvikoski, T., and Sulkava, R. (2010). Education, the Brain and Dementia: Neuroprotection or Compensation? EClipSE Collaborative Members. Brain, 133(8), 2210–2216.

Breyer, F., Lorenz, N., and Niebel, T. (2015). Health Care Expenditures and Longevity: Is There a Eubie Blake Effect? The European journal of health economics, 16(1), 95–112.

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

Britton, J., and Propper, C. (2016). Teacher Pay and School Productivity: Exploiting Wage Regulation. Journal of Public Economics, 133(Supplement C), 75 – 89.

Brookhart, M., Schneeweiss, S., Rothman, K., Glynn, R., Avorn, J., and Stürmer, T. (2006). Variable Selection for Propensity Score Models. American Journal of Epidemiology, 163(12), 1149–1156.

Budria, S., and Ferrer-i Carbonell, A. (2012). Income Comparisons and Non-cognitive Skills. SOEPpapers No. 441, (pp. 1–29).

Cai, L. (2010). The Relationship Between Health and Labour Force Participation: Evidence from a Panel Data Simultaneous Equation Model. Labour Economics, 17(1), 77 – 90.

Card, D. (1995). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. In L. Christofides, K. Grant, and S. R. (Eds.) Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp, (pp. 201–222). University of Toronto Press.

Card, D. (2001). Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. Econometrica, 69(5), 1127–1160.

Carmichael, F., and Charles, S. (1998). The Labour Market Costs of Community Care. Journal of Health Economics, 17(6), 747–765.

Carmichael, F., and Charles, S. (2003). The Opportunity Costs of Informal Care: Does Gender Matter? Journal of Health Economics, 22(5), 781–803.

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.

Casado-Marín, D., Pilar García-Gómez, P., and Ángel López-Nicolás (2011). Informal Care and Labour Force Participation among Middle-aged Women in Spain. SERIEs, 2(1), 1–29.

Case, A., and Paxson, C. (2009). Early Life Health and Cognitive Function in Old Age. The American Economic Review, 99(2), 104.

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.

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.

Ciani, E. (2012). Informal Aadult Care and Caregivers' Employment in Europe. Labour Economics, 19(2), 155–164.

Clark, D., and Martorell, P. (2014). The Signaling Value of a High School Diploma. Journal of Political Economy, 122(2), 282–318.

Clark, G. (2012). What is the True Rate of Social Mmobility in Sweden? A Surname Analysis, 1700-2012. Manuscript, Univ. California, Davis.

Classen, T. J. (2010). Measures of the Intergenerational Transmission of Body Mass Iindex Between Mothers and their Children in the United States, 1981–2004. Economics & Human Biology, 8(1), 30–43.

Cobb-Clark, D. A., and Schurer, S. (2013). Two Economists' Musings on the Stability of Locus of Control. The Economic Journal, 123(570), F358–F400.

Coe, N. B., and van Houtven, C. H. (2009). Caring for Mom and Neglecting Yourself? The Health Effects of Caring for an Elderly Parent. Health Economics, 18(9), 991–1010.

Colvez, A., Joel, M.-E., Ponton-Sanchez, A., and Royer, A.-C. (2002). Health Status and Work Burden of Alzheimer Patients' Informal Caregivers: Comparisons of Five Different Care Programs in the European Union. Health Policy, 60(3), 219,233.

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.

Costa-Font, J., and Gil, J. (2013). Intergenerational and Socioeconomic Gradients of Child Obesity. Social Science & Medicine, (93), 29–37.

Coutinho, R., David, R. J., and Collins, J. W. (1997). Relation of Parental Birth Weights to Infant Birth Weight among African Americans and Whites in Illinois: A Transgenerational Study. American Journal of Epidemiology, 146(10), 804–809.

Crespo, L., and Mira, P. (2014). Caregiving to Elderly Parents and Employment Status of European Mature Women. The Review of Economics and Statistics, 96(3), 693–709.

Crimmins, E. M., and Finch, C. E. (2006). Infection, Inflammation, Height, and Longevity. Proceedings of the National Academy of Sciences of the United States of America, 103(2), 498–503.

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

Cunha, F., Heckman, J. J., Lochner, L. J., and Masterov, D. V. (2006). Interpreting the Evidence on Life Cycle Skill Formation. In E. A. Hanushek, and F. Welch (Eds.) Handbook of the Economics of Education, vol. 1 of Handbook of the Economics of Education, chap. 12. North-Holland.

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 Moretti, E. (2007). Biology as Destiny? Short– and Long–Run Determinants of Intergenerational Transmission of Birth Weight. Journal of Labor Economics, 25(2), 231–264.

Cutler, D. M., Deaton, A., and Lleras-Muney, A. (2006). The Determinants of Mortality. Journal of Economic Perspectives, 20(3), 97–120.

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.

Datar, A., Kilburn, R., and Loughran, D. S. (2010). Endowments and Parental Investments in Infancy and Early Childhood. Demography, 47(1), 145–162.

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.

Dehne, M., and Schupp, J. (2007). Persönlichkeitsmerkmale im Sozio-ökonomischen Panel (SOEP) - Konzept, Umsetzung und empirische Eigenschaften. Tech. rep., DIW Berlin.

Der Spiegel (1967). Doktortitel Nach Sechs Semestern? Der Spiegel, Oktober 9, 1967(42), 54–62.

Di Novi, C., Jacobs, R., and Migheli, M. (2013). The Quality of Life of Female Informal Caregivers: From Scandinavia to the Mediterranean Sea. CHE Research Paper 84,

Centre for Health Economics, University of York.

Die Zeit (1967). Warenhaus der Ausbildung. Die Zeit, August 4, 1967(31), 28.

Do, Y. K., Norton, E. C., Stearns, S., and Houtven, C. H. V. (2014). Informal Care and Caregiver's Health. Health Economics, forthcoming.

Doblhammer, G., van den Berg, G. J., and Lumey, L. H. (2011). Long–term Effects of Famine on Life Expectancy: A Re–analysis of the Great Finnish Famine of 1866–1868. IZA Discussion Papers, 5534.

Drake, A. (2004). The Intergenerational Effects of Fetal Programming: Non-genomic Mechanisms for the Inheritance of Low Birth Weight and Cardiovascular Risk. Journal of Endocrinology, 180(1), 1–16.

Duflo, E., Dupas, P., and Kremer, M. (2015). Education, HIV, and Early Fertility: Experimental Evidence from Kenya. American Economic Review, 105(9), 2757–2797.

Dunkin, J. J., and Anderson-Hanley, C. (1998). Dementia Caregiver Burden - A Review of the Literature and Guidelines for Assessment and Intervention. Neurology, 51(1), 53–60.

Durchhardt, C., and Gerdes, A. (2012). NEPS Technical Report for Mathematics – Scaling Results of Starting Cohort 3 in Fifth Grade. Tech. rep., NEPS Working Papers No.19.

Edvinsson, S. (2000). The Demographic Data Base at Umeå University: A resource for historical studies. Handbook of International Historical Microdata for Population Research, (pp. 231–248).

Edvinsson, S., Gardharsdóttir, Ö., and Thorvaldsen, G. (2008). Infant Mortality in the Nordic Countries, 1780–1930. Continuity and Change, 23(03), 457–485.

Ettner, S. L. (1995). The Impact of "Parent Care" on Female Labor Supply Decisions. Demography, 32(1), 63–80.

Ettner, S. L. (1996). The Opportunity Costs of Elder Care. The Journal of Human Resources, 31(1), 189–205.

European Union, Committee of the Regions. (2006). The Impact of Demographic Change on European Regions. Tech. rep.

Federation of Swedish Genealogical Societies (2009). Buried in Sweden (Begravda I Sverige).

Fevang, E., Kverndokk, S., and Roed, K. (2012). Labor Supply in the Terminal Stages of Lone Parents' Lives. Journal of Population Economics, 25(4), 1399–1422.

Figlio, D., Guryan, J., Karbownik, K., and Roth, J. (2014). The Effects of Poor Neonatal Health on Children's Cognitive Development. American Economic Review, 104(12), 3921–55.

Figlio, D. N. (1997). Teacher Salaries and Teacher Quality. Economics Letters, 55(2), 267–271.

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.

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.

Franzmann, G. (2006). The development of the educational system in germany: Comprehensive schools 1960–2000.

Franzmann, G. (2015). Histat-Datenkompilation: BevÃ¶lkerung nach Alter in Jahren und nach Geschlecht für das Deutsche Reich, die frühere Bundesrepublik und Deutschland, 1871-2010.

Gallicchio, L., Siddiqi, N., Langenberg, P., and Baumgarten, M. (2002). Gender Differences in Burden and Depression among Informal Caregivers of Demented Elders in the Community. International Journal of Geriatric Psychiatry, 17(2), 154–163.

García-Gómez, P. (2011). Institutions, Health Shocks and Labour Market Outcomes across Europe. Journal of Health Economics, 30(1), 200–213.

García-Gómez, P., van Kippersluis, H., O'Donnell, O., and van Doorslaer, E. (2013). Long-Term and Spillover Effects of Health Shocks on Employment and Income. Journal of Human Resources, 48(4), 873–909.

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.

Gehrer, K., Zimmermann, S., Artelt, C., and Weinert, S. (2012). The Assessment of Reading Competence (Including Sample Items For Grade 5 and 9). Tech. rep., NEPS research data – Leibnitz Institute for Educational Trajectories.

Gerdtham, U.-G., and Ruhm, C. J. (2006). Deaths Rise in Good Economic Times: Evidence from the OECD. Economics & Human Biology, 4(3), 298–316.

German Federal Statistical Office (2016). Endgültige durchschnittliche Kinderzahl der Frauenkohorten. Tech. rep., German Federal Statistical Office (Statistisches Bundesamt), Wiesbaden.

German Federal Statistical Office (various issues, 1959–1991). Statistisches Jahrbuch für die Bundesrepublik Deutschland. Tech. rep., 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.

Geyer, J., and Korfhage, T. (2015a). Long-term Care Insurance and Carers' Labor Supply – A Structural Model. Health Economics, 24(9), 1178–1191.

Geyer, J., and Korfhage, T. (2015b). Long-Term Care Reform and the Labor Supply of Household Members: Evidence from a Quasi-Experiment. SOEPpapers on Multidisciplinary Panel Data Research 785, DIW Berlin, The German Socio-Economic Panel (SOEP).

Gill, S. C., Butterworth, P., Rodgers, B., and Mackinnon, A. (2007). Validity of the Mental Health Component Scale of the 12-item Short-Form Health Survey (MCS-12) as Measure of Common Mental Disorders in the General Population. Psychiatry Research, 152(1), 63 – 71.

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.

Goldenberg, R. L., Culhane, J. F., Iams, J. D., and Romero, R. (2008). Epidemiology and Causes of Preterm Birth. The Lancet, 371(9606), 75–84.

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.

Goldstein, I. (2013). Chapter 36 - Empirical Literature on Financial Crises: Fundamentals vs. Panic. In G. C. B. C. L. Schmukler (Ed.) The Evidence and Impact of Financial Globalization, (pp. 523 – 534). San Diego: Academic Press.

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. (2006). Education and Nonmarket Outcomes. Handbook of the Economics of Education, 1, 577–633.

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.

Haider, S., and Solon, G. (2006). Life-Cycle Variation in the Association between Current and Lifetime Earnings. American Economic Review, 96(4), 1308–1320.

Hales, C. N., and Barker, D. J. P. (1992). Type 2 (Non-insulin-dependent) Diabetes Mellitus: The Thrifty Phenotype Hypothesis. Diabetologia, 35(7), 595–601.

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. (1971). Teacher Characteristics and Gains in Student Achievement: Estimation using Micro Data. American Economic Review, 61(2), 280–288.

Hanushek, E. A., and Rivkin, S. G. (2006). Teacher Quality. Handbook of the Economics of Education, 2, 1051–1078.

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

Harper, S. (2014). Economic and Social Implications of Aging societies. Science, 346(6209), 587–591.

Hatton, T. J. (2011). Infant mortality and the Health of Survivors: Britain, 1910–50. The Economic History Review, 64(3), 951–972.

Heckman, J. J. (1990). Varieties of Selection Bias. The American Economic Review, 80(2), pp. 313–318.

Heckman, J. J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. Econometrica, 66(5), p 1017–1098.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., and Yavit, A. (2010). The Rate of Return to the HighScope Perry Preschool Program. Journal of Public Economics, 94(1-2), 114 – 128.

Heckman, J. 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), 2052–86.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006a). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. Journal of Labor Economics, 24(3), 411–482.

Heckman, J. J., Urzua, S., and Vytlacil, E. J. (2006b). 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 J. J. Heckman, and E. E. Leamer (Eds.) Handbook of Econometrics, vol. 6 of Handbook of Econometrics, chap. 71. Elsevier.

Heger, D. (2014). Work and Well-Being of Informal Caregivers in Europe. Ruhr Economic Papers 0512, Rheinisch-Westfälisches Institut für Wirtschaftsforschung,

Ruhr-Universität Bochum, Universität Dortmund, Universität Duisburg-Essen.

Heger, D., and Korfhage, T. (2017). Does the Negative Effect of Caregiving on Work Persist Over Time? Tech. Rep. 703, Ruhr Economic Papers.

Heitmueller, A. (2007). The Chicken or the Egg?: Endogeneity in Labour Market Participation of Informal Carers in England. Journal of Health Economics, 26(3), 536 – 559.

Heitmueller, A., and Inglis, K. (2007). The earnings of informal carers: Wage differentials and opportunity costs. Journal of Health Economics, 26(4), 821–841.

Hertz, T., Jayasundera, T., Piraino, P., Selcuk, S., Smith, N., and Verashchagina, A. (2007). The Inheritance of Educational Inequality: International Comparisons and Fifty-year Trends. The BE Journal of Economic Analysis & Policy, 7(2).

Ho, S. C., Chan, A., Woo, J., Chong, P., and Sham, A. (2009). Impact of Caregiving on Health and Quality of Life: A Comparative Population-Based Study of Caregivers for Elderly Persons and Noncaregivers. The journals of gerontology. Series A, biological sciences and medical sciences, 64(8), 873–879.

Hochberg, Z., Feil, R., Constancia, M., Fraga, M., Junien, C., Carel, J., Boileau, P., Le Bouc, Y., Deal, C., Lillycrop, K., Scharfmann, R., Sheppard, A., Skinner, M., Szyf, M., Waterland, R., Waxman, D., Whitelaw, E., Ong, K., and Albertsson-Wikland, K. (2011). Child Health, Developmental Plasticity, and Epigenetic Programming. Endocr Rev, 32(2), 159–224.

Holford, T. R., Meza, R., Warner, K. E., Meernik, C., Jeon, J., Moolgavkar, S. H., and Levy, D. T. (2014). Tobacco Control and the Reduction in Smoking-related Premature Deaths in the United States, 1964-2012. Jama, 311(2), 164–171.

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, 305–327.

Imbens, G. W. (2003). Sensitivity to Exogeneity Assumption in Program Evaluation. American Economic Review, 93(2), 126–132.

Imbens, G. W., and Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. Econometrica, 62(2), 467–475.

Jackson, C. K., Rockoff, J. E., and Staiger, D. O. (2014). Teacher Effects and Teacher-Related Policies. Annual Review of Economics, 6(1), 801–825.

Jantti, M., Bratsberg, B., Roed, K., Raaum, O., Naylor, R., Osterbacka, E., Bjorklund, A., and Eriksson, T. (2006). American Exceptionalism in a New Light: A Comparison of Intergenerational Earnings Mobility in the Nordic Countries, the United Kingdom and the United States. IZA discussion paper.

Jirtle, R. L., and Skinner, M. K. (2007). Environmental Epigenomics and Disease Susceptibility. Nature reviews genetics, 8(4), 253–262.

Johnson, R. W., and Sasso, A. T. L. (2000). The trade-off between hours of paid employment and time assistance to elderly parents at midlife. Working paper, The Urban Institute.

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.

Kamhöfer, D. A., and Schmitz, H. (2016). Reanalyzing Zero Returns to Education in Germany. Journal of Applied Econometrics, 31(5), 912–919. Jae.2461.

Kamhöfer, D. A., and Schmitz, H. (2016). Reanalyzing Zero Returns to Education in Germany. Journal of Applied Econometrics, 31(5), 912–919.

Kamhöfer, D. A., Schmitz, H., and Westphal, M. (2015). Heterogeneity in Marginal Non-monetary Returns to Higher Education. RWI Essen 591, Ruhr Economic Papers.

Kamhöfer, D. A., Schmitz, H., and Westphal, M. (2017). Heterogeneity in Marginal Non-monetary Returns to Higher Education. Journal of the European Economic Association, forthcoming.

Kamhöfer, D. A., and Westphal, M. (2017). Fertility Effects of College Education: Evidence from the German Educational Expansion. Ruhr Economic Papers 717, RWI Essen.

Karlsson, M., and Klohn, F. (2014). Testing the Red Herring Hypothesis on an Aggregated Level: Ageing, Time-to-death and Care Costs for Older People in Sweden. The European Journal of Health Economics, 15(5), 533–551.

Karlsson, M., Nilsson, T., and Pichler, S. (2014). The Impact of the 1918 Spanish Flu Epidemic on Economic Performance in Sweden: An Investigation into the Consequences of an Extraordinary Mortality Shock. Journal of health economics, 36, 1–19.

Khot, U. N., Khot, M. B., Bajzer, C. T., Sapp, S. K., Ohman, E. M., Brener, S. J., Ellis, S. G., Lincoff, A. M., and Topol, E. J. (2003). Prevalence of Conventional Risk Factors in Patients With Coronary Heart Disease. Jama, 290(7), 898–904.

Kim, Y., Sikoki, B., Strauss, J., and Witoelar, F. (2014). Intergenerational Correlations of Health Among Older Adults: Empirical Evidence from Indonesia. The Journal of the Economics of Ageing, (0), –.

Kobrak, C., and Mira, W. e. (Eds.) (2013). History and Financial Crises: Lessons from the 20th Century.. Routledge, New York.

Koedel, C. (2009). An Empirical Analysis of Teacher Spillover Effects in Secondary School. Economics of Education Review, 28(6), 682 – 692.

Köhler, H., and Lundgreen, P. (2015). General Secondary Schools in the Federal Republic of Germany from 1949 to 2010.

Lakdawalla, D. (2001). The Declining Quality of Teachers. Tech. rep., National Bureau of Economic Research.

Lang, F., Weiss, D., Stocker, A., and von Rosenbladt, B. (2007). 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.

Langley Evans, S. (2015). Nutrition in Early Life and the Programming of Adult Disease: A Review. Journal of Human Nutrition and Dietetics, 28, 1–14.

Lawson, N., and Spears, D. (2014). What Doesn't Kill You Makes You Poorer: Adult Wages and the Early-life Disease Environment in India. World Bank Policy Research Working Paper, (7121).

Lechner, M. (2008). Matching Estimation of Dynamic Treatment Models: Some Practical Issues. In D. Millimet, J. Smith, and V. E. (Eds.) Modelling and Evaluating Treatment Effects in Econometrics, vol. 21 of Advances in Econometrics, (pp. 289–333).

Lechner, M. (2009a). Long-run Labour Market and Health Effects of Individual Sport Activities. Journal of Health Economics, 28, 839–854.

Lechner, M. (2009b). Sequential Causal Models for the Evaluation of Labor Market Programs. Journal of Business & Economic Statistics, 27, 71–83.

Lechner, M., and Miquel, R. (2010). Identification of the Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions. Empirical Economics, 39(1), 111–137.

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.

Lee, S., Colditz, G. A., Berkman, L. F., and Kawachi, I. (2003). Caregiving and Risk of Coronary Heart Disease in U.S. Women: A Prospective Study. American Journal of Preventive Medicine, 24(2), 113–119.

Leigh, A. (2010). Informal Care and Labor Market Participation. Labour Economics, 17(1), 140–149.

Lengerer, A., Schroedter, J., Boehle, M., Hubert, T., and Wolf, C. (2008). Harmonisierung der Mikrozensen 1962 bis 2005. Tech. rep., GESIS-Methodenbericht 12/2008, GESIS–Leibniz Institute for the Social Sciences, German Microdata Lab, Mannheim.

LIfBi (2011). Starting Cohort 6 Main Study 2010/11 (B67) Adults Information on the Competence Test. Tech. rep., Leibniz Institute for Educational Trajectories (LIfBi) – National Educational Panel Study.

LIfBi (2015). Startkohorte 6: Erwachsene (SC6) – Studienübersicht Wellen 1 bis 5. Tech. rep., Leibniz Institute for Educational Trajectories (LIfBi) – National Educational Panel Study.

Lindahl, M., Palme, M., Sandgren Massih, S., and Sjögren, A. (2012). The Intergenerational Persistence of Human Capital: An Empirical Analysis of Four Generations.

Lindeboom, M., Portrait, F., and Van den Berg, G. J. (2010). Long-run Effects on Longevity of a Nutritional Shock Early in Life: The Dutch Potato Famine of 1846–1847. Journal of health economics, 29(5), 617–629.

Lipszyc, B., Sail, E., Xavier, A., et al. (2012). Long-term care: Need, Use and Expenditure in the EU-27. Tech. rep., Directorate General Economic and Financial Affairs (DG ECFIN), European Commission.

Loeb, S., and Page, M. E. (2000). Examining the Link between Teacher Wages and Student Outcomes: The Importance of Alternative Labor Market Opportunities and Non-pecuniary Variation. The Review of Economics and Statistics, 82(3), 393–408.

Lundgreen, P., and Schallmann, J. (2013). Datenhandbuch zur deutschen Bildungsgeschichte, vol. XI: Die Lehrer an den Schulen in der Bundesrepublik Deutschland 1949-2009. Vandenhoeck & Ruprecht.

Lundgreen, P., and Schwibbe, G. (2008). Berufliche Schulen und Hochschulen in der Bundesrepublik Deutschland 1949-2001 Teil II: Hochschulen. Tech. Rep. Deutschland ZA8202 Datenfile Version 1.0.0.

Lundh, C. (1999). The Social Mobility of Servants in Rural Sweden, 1740–1894. Continuity and Change, 14(01), 57–89.

Lundh, C., and Prado, S. (2015). Markets and Politics: The Swedish Urban–rural Wage Gap, 1865–1985. European Review of Economic History, 19(1), 67–87.

Magnusson, L. (2010). Sveriges ekonomiska historia.

Manski, C. F. (1995). Identification Problems in the Social Sciences. Cambridge and Mass: Harvard University Press.

Marcus, J. (2013). The Effect of Unemployment on the Mental Health of Spouses – Evidence from Plant Closures in Germany. Journal of Health Economics, 32, 546–558.

Marcus, J. (2014). Does Job Loss Make You Smoke and Gain Weight? Economica, 324(81), 626–648.

Margerison-Zilko, C., Catalano, R., Hubbard, A., and Ahern, J. (2011). Maternal Exposure to Unexpected Economic Contraction and Birth Weight for Gestational Age. Epidemiology, 22(6), 855–858.

Max Planck Institute for Demographic Research, G., and Vienna Institute of Demography, A. (2014). Human Fertility Database. Tech. rep.

Mazumder, B. (2005). Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data. Review of Economics and Statistics, 87(2), 235–255.

Mazumder, B. (2008). Does Education Improve Health? A Reexamination of the Evidence from Compulsory Schooling Laws. Economic Perspectives, (Q II), 2–16.

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.

McRae, R. R., and John, O. P. (1992). An Introduction to the Five Factor Model and Its Applications. Journal of Personality and Social Psychology, 60(2), 175–215.

Medicinalstyrelsen (1934). Socialhygienisk undersökning i Västerbottens och Norrbottens län 1929-1931. Lund, Sweden, Håkan Olssons Boktryckeri.

Meng, A. (2012). Informal Caregiving and the Retirement Decision. German Economic Review, 13(3), 307–330.

Meng, A. (2013). Informal Home Care and Labor-force Participation of Household Members. Emprical Economics, 44(2), 959–979.

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.

Michaud, P.-C., Heitmueller, A., and Nazarov, Z. (2010). A Dynamic Analysis of Informal Care and Employment in England. Labour Economics, 17(3), 455–465.

Mitchell, B. R. (Ed.) (1998). International Historical Statistics: Europe, 1750-1993. London: Macmillan Reference; New York: Stockton Press.

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.

Mor, G., and Cardenas, I. (2010). The Immune System in Pregnancy: A Unique Complexity. American Journal of Reproductive Immunology, 63(6), 425–433.

Müller, R., Unger, R., and Rothgang, H. (2010). Reicht eine zweijährige Familien-Pflegezeit für Arbeitnehmer? Wie lange Angehörige zu Hause gepflegt werden. Soziale Sicherheit. Zeitschrift für Arbeit und Soziales, 10(6–7), 230–237.

Nagler, M., Piopiunik, M., and West, M. R. (2015). Weak Markets, Strong Teachers: Recession at Career Start and Teacher Effectiveness. NBER Working Papers 21393, National Bureau of Economic Research, Inc.

Nannicini, T. (2007). Simulation-based Sensitivity Analysis for Matching Estimators. Stata Journal, 7(3), 334–350.

Neumayer, E. (2004). Recessions Lower (Some) Mortality Rates: Evidence from Germany. Social Science & Medicine, 58(6), 1037–1047.

NRW (1971a). Sachstandsbericht des Ministers für Wissenschaft und Forschung. Report, Ministry of Science and Research of the state of North Rhine-Westphalia (NRW), March 2, 1971, Düsseldorf.

NRW (1971b). Stellungnahme der Staatskanzlei zum Entwurf der Kabinettvorlage des Ministers für Wissenschaft und Forschung. Report, Office of the Prime Minister of the state of North Rhine-Westphalia (NRW), April 19, 1971, Düsseldorf.

NRW (1971c). Schreiben des Ministers für Wissenschaft und Forschung an die Staatskanzlei. Report, Ministry of Science and Research of the state of North Rhine-Westphalia (NRW), May 24, 1971, Düsseldorf.

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

Oddens, B., Vemer, H., Visser, A., and Ketting, E. (1993). Contraception in Germany: A Review. Advances in Contraception, 9, 105–116.

Odin, M. (1934). Sjukdomar och sjukdomsfrekvens i övre norrland särskilt med hänsyn til födans sammansättning (diseases and disease incidence in upper norrland with particular focus on the dietary composition). Socialhygienisk undersökning i Västerbottens och Norrbottens län. Lund, Sweden: Håkan Olssons Boktryckeri.

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).

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. G. (2011). Priceless: The Nonpecuniary Benefits of Schooling. Journal of Economic Perspectives, 25(1), 159–84.

Osterroth, F., and Schuster, D. (2000). Chronik der deutschen sozialdemokratie.

Pearce, J., and Langley Evans, S. (2013). The Types of Food Introduced During Complementary Feeding and Risk of Childhood Obesity: A Systematic Review. International Journal of Obesity, 37, 477–485.

Pei, Z., Pischke, J.-S., and Schwandt, H. (2017). Poorly Measured Confounders are More Useful on the Left Than on the Right. Tech. rep., National Bureau of Economic Research.

Pelkowski, J. M., and Berger, M. C. (2004). The Impact of Health on Employment, Wages, and Hours Worked Over the Life Cycle. The Quarterly Review of Economics and Finance, 44, 102–121.

Pettersson-Lidbom, P. (2014). Midwives and Maternal Mortality: Evidence from a Midwifery Policy Experiment in Sweden in the 19th Century. Tech. rep., mimeo.

Picht, G. (1964). Die deutsche Bildungskatastrophe: Analyse und Dokumentation. Walter Verlag.

Picht, G. (1965). Die deutsche Bildungskatastrophe. Deutscher Taschenbuch Verlag München.

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.

Pohl, S., and Carstensen, C. H. (2012). NEPS Technical Report for Mathematics – Scaling the Data of the Competence Tests. Tech. rep., NEPS Working Papers No.19.

Prentice, A. M. (2006). The Emerging Epidemic of Obesity in Developing Countries. International Journal of Epidemiology, 35(1), 93–99.

Raute, A. (2017). Can Financial Incentives Reduce the Baby Gap? Evidence from a Reform in Maternity Leave Benefits. NBER Working Papers 23793, National Bureau of Economic Research.

Reichert, A., and Tauchmann, H. (2011). The Causal Impact of Fear of Unemployment on Psychological Health. Ruhr Economic Papers 266 266, Essen.

Richter, A., and Robling, P. O. (2013). Multigenerational Effects of the 1918-19 Influenza Pandemic in Sweden. Swedish Institute for Social Research, (5).

Riphahn, R. T., and Wiynck, F. (2017). Fertility Effects of Child Benefits. Journal of Population Economics, forthcoming.

Robinson, P. M. (1988). Root- N-Consistent Semiparametric Regression. Econometrica, 56(4), 931–954.

Rockoff, J. E., Jacob, B. A., Kane, T. J., and Staiger, D. O. (2011). Can You Recognize an Effective Teacher When You Recruit One? Education Finance and Policy, 6(1), 43–74.

Rohwedder, S., and Willis, R. J. (2010). Mental Retirement. Journal of Economic Perspectives, 24(1), 119–38.

Rosenbaum, P. R., and Rubin, D. B. (1983). Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. Journal of the Royal Statistical Society. Series B (Methodological), 45(2), 212–218.

Rothgang, H. (2010). Social Insurance for Long-term Care: An Evaluation of the German Model. Social Policy & Administration, 44(4), 436–460.

Roy, A. D. (1951). Some Thoughts on the Distribution of Earnings. Oxford economic papers, 3(2), 135–146.

Royer, H. (2009). Separated at Girth: U.S. Twin Estimates of the Effects of Birth Weight. American Economic Journal: Applied Economics, 1(1), 49–85.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Journal of Educational Psychology, 56(5), 688–701.

Rubin, D. B. (1979). Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. Journal of the American Statistical Association, 74(366), pp. 318–328.

Ruhm, C. J. (2000). Are Recessions Good for Your Health? The Quarterly Journal of Economics, 115(2), 617–650.

Ruhm, C. J. (2003). Good Times Make You Sick. Journal of Health Economics, 22(4), 637–658.

Ruhm, C. J. (2004). Macroeconomic Conditions, Health and Mortality. Tech. rep., National Bureau of Economic Research.

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.

Salyers, M. P., Bosworth, H. B., Swanson, J. W., Lamb-Pagone, J., and Osher, F. C. (2000). Reliability and Validity of the SF-12 Health Survey Among People with Severe Mental Illness. Medical care, 38(11), 1141–1150.

Schmidt, I. M., Jørgensen, M., and Michaelsen, K. F. (1995). Height of Conscripts in Europe: Is Postneonatal Mortality a Predictor? Annals of Human Biology, 22(1), 57–67.

Schmidt, M., and Schneekloth, U. (2011). Abschlussbericht zur Studie "Wirkungen des Pflege-Weiterentwicklungsgesetzes".

Schmitz, H. (2011). Why are the Unemployed in Worse Health? The Causal Effect of Unemployment on Health. Labour Economics, 18(1), 71–78.

Schmitz, H., and Stroka, M. A. (2013). Health and the Double Burden of Full-Time Work and Informal Care Provision: Evidence From Administrative Data. Labour Economics, 24, 305–322.

Schmitz, H., and Westphal, M. (2015). Short- and Medium-term Effects of Informal Care Provision on Female Caregivers' Health. Journal of Health Economics, 42(C), 174–185.

Schmitz, H., and Westphal, M. (2017). Informal Care and Long-term Labor Market Outcomes. Journal of Health Economics, 56(Supplement C), 1 – 18.

Schneekloth, U., and Leven, I. (2003). Hilfe und Pflegebedürftige in Privathaushalten in Deutschland 2002.

Schneeweis, N., Skirbekk, V., and Winter-Ebmer, R. (2014). Does Education Improve Cognitive Performance Four Decades After School Completion? Demography, 51(2), 619–643.

Schnittjer, I., and Duchhardt, C. (2015). Mathematical Competence: Framework and Exemplary Test Items. Tech. rep., NEPS research data – Leibnitz Institute for Educational Trajectories.

Schofer, E., and Meyer, J. W. (2005). The worldwide expansion of higher education in the twentieth century. American sociological review, 70(6), 898–920.

Scholte, R., van den Berg, G. J., and Lindeboom, M. (2012). Long-Run Effects of Gestation During the Dutch Hunger Winter Famine on Labor Market and Hospitalization Outcomes.

Schulz, E. (2010). The Long-Term Care System for the Elderly in Germany. DIW Discussion Paper 1039, DIW Berlin.

Schulz, R., O'Brien, A., Bookwala, J., and Fleissner, K. (1995). Psychiatric and Physical Morbidity Effects of Dementia Caregiving: Prevalence, Correlates, and Causes. Gerontologist, 35(6), 771–791.

Shaw, W. S., Patterson, T. L., Ziegler, M. G., Dimsdale, J. E., Semple, S. J., and Grant, I. (1999). Accelerated Risk of Hypertensive Blood Pressure Recordings among Alzheimer Caregivers. Journal of Psychosomatic Research, 46(3), 215–227.

Siegler, B. (2012). The Effect of University Openings on Local Human Capital Formation: Difference-in-Differences Evidence from Germany. Tech. rep., BGPE Discussion Paper.

Skatteverket (2015). Sveriges församlingar genom tiderna:församlingar med förlossningsanstalter fram till 1947, http://www.skatteverket.se/privat/ folkbokforing/omfolkbokforing.html.

Skira, M. M. (2015). Dynamic Wage and Employment Effects of Elder Parent Care. International Economic Review, 56(1), 63–93.

Sköld, P. (1996). From Inoculation to Vaccination: Smallpox in Sweden in the Eighteenth and Nineteenth Centuries. Population Studies, 50(2), 247–262.

Skopek, J., Pink, S., and Bela, D. (2012). Starting Cohort 3: 5th Grade (SC3) – SUF Version 1.0.0 Data Manual. Tech. rep., NEPS research data – Leibnitz Institute for Educational Trajectories.

Sobotka, T. (2004). Postponement of Childbearing and Low Fertility in Europe. Dutch University Press Amsterdam.

Socialstyrelsen, S. (1938). Levnadsvillkor och hushållsvanor i städer och industriorter omkring år 1933. Sveriges officiella statistik.

SOEP Group (2014). SOEP 2013 – Documentation of Person-related Status and Generated Variables in PGEN for SOEP v30.

Sonnenberg, B., Riediger, M., Wrzus, C., and Wagner, G. G. (2012). Measuring Time Use in Surveys - How Valid are Time Use Questions in Surveys? Concordance of Survey and Experience Sampling Measures. SOEP Discussion Paper 390, (390).

Statistics Sweden (1935). Årsbok för Sveriges kommuner: Statistical yearbook of administrative districts of Sweden, 1931–1935.

Statistics Sweden (1997). Dödsorsaker 1995.

Statistics Sweden (2005). Multi-generation Register 2005 – A Description of Contents and Quality. Örebro: Statistics Sweden, (pp. 1–88).

Statistisches Bundesamt (2014). Bildungsstand der Bevölkerung.

Statistisches Bundesamt (2015). Pflegestatistik 2013. Pflege im Rahmen der Pflegeversicherung. Deutschlandergebnisse.

Stephen, M. A., Townsend, A. L., Martire, L. M., and Druley, J. A. (2001). Balancing Parent Care with other Roles: Interrole Conflict of adult Daughter Caregivers. The journals of gerontology. Series B, psychological sciences and social sciences, 56(1), P24–P31.

Stephens, M. J., and Yang, D.-Y. (2014). Compulsory Education and the Benefits of Schooling. American Economic Review, 104(6), 1777–92.

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.

Tennstedt, S., Cafferata, G. L., and Sullivan, L. (1992). Depression among Caregivers of Impaired Elders. Journal of ageing and health, 4(1), 58–76.

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).

The National Board of Health & Welfare (2014). Dödsorsaksregistret.

Thompson, O. (2014). Genetic Mechanisms in the Intergenerational Transmission of Health. Journal of Health Economics, 35, 132–146.

Torell, U. (2013). När sockret blev vardagsmat. Tidsskrift for kulturforskning, 12(3-4).

Trannoy, A., Tubeuf, S., Jusot, F., and Devaux, M. (2010). Inequality of Opportunities in Health in France: A First Pass. Health Economics, 19(8), 921–938.

Van den Berg, B., and Ferrer-i Carbonell, A. (2007). Monetary Valuation of Informal Care: The Well-being Valuation Method. Health Economics, 16(11), 1227–1244.

Van den Berg, B., Fiebig, D. G., and Hall, J. (2014). Well-being Losses Due to Caregiving. Journal of Health Economics, 35(0), 123 – 131.

Van den Berg, B., and Spauwen, P. (2006). Measurement of Informal Care: An Empirical Study into the Valid Measurement of Time Spent on Informal Caregiving. Health Economics, 15(5), 447–460.

Van den Berg, G. J., Doblhammer, G., and Christensen, K. (2009). Exogenous Determinants of Early-life Conditions, and Mortality Later in Life. Social Science & Medicine, 68(9), 1591–1598.

Van den Berg, G. J., Doblhammer-Reiter, G., and Christensen, K. (2011). Being Born Under Adverse Economic Conditions Leads to a Higher Cardiovascular Mortality Rate Later in Life: Evidence Based on Individuals Born at Different Stages of the Business Cycle. Demography, 48(2), 507–530.

Van den Berg, G. J., Lindeboom, M., and Portrait, F. (2006). Economic Conditions Early in Life and Individual Mortality. The American Economic Review, (pp. 290–302).

Van den Berg, G. J., and Pinger, P. R. (2014). Transgenerational Effects of Childhood Conditions on Third Generation Health and Education Outcomes. SOEP Paper, (709), 343–356.

Van Houtven, C., Wilson, M., and Clipp, E. (2005). Informal Care Intensity and Caregiver Drug Utilization. Review of Economics of the Household, 3(4), 415–433.

Van Houtven, C. H., Coe, N. B., and Skira, M. M. (2013). The Effect of Informal Care on Work and Wages. Journal of Health Economics, 32(1), 240–252.

Vilagut, G., Forero, C. G., Pinto-Meza, A., Haro, J. M., de Graaf, R., Bruffaerts, R., Kovess, V., de Girolamo, G., Matschinger, H., Ferrer, M., and Alonso, J. (2013). The Mental Component of the Short-Form 12 Health Survey (SF-12) as a Measure of Depressive Disorders in the General Population: Results with Three Alternative Scoring Methods. Value in Health, 16(4), 564 – 573.

Vytlacil, E. (2002). Independence, Monotonicity, and Latent Index Models: An Equivalence Result. Econometrica, 70(1), 331–341.

Wagner, G. G., Frick, J. R., and Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP): Scope, Evolution, and Enhancements. Journal of Applied Social Science Studies (Schmollers Jahrbuch: Zeitschrift für Wirtschafts- und Sozialwissenschaften), 127(1), 139–169.

Ware, J. E., Kosinski, M., and Keller, S. D. (1996). A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary Tests of Reliability and Validity. Medical care, 34(3), 220–233.

Weber, M., Grote, V., Closa-Monasterolo, R., Escribano, J., Langhendries, J., Dain, E., Giovannini, M., Verduci, E., Gruszfeld, D., Socha, P., and Koletzko, B. (2014). Lower Protein Content in Infant Formula Reduces BMI and Obesity Risk at School Age: Follow-up of a Randomized Trial. American Journal of Clinical Nutrition, 99(5), 1041–1051.

Weinehall, L., Westman, G., Hellsten, G., Boman, K., Hallmans, G., Pearson, T. A., and Wall, S. (1999). Shifting the Distribution of Risk: Results of a Community Intervention in a Swedish Programme for the Prevention of Cardiovascular Disease. Journal of Epidemiology and Community Health, 53(4), 243–250.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., and Carstensen, C. (2011). Development of Competencies Across the Life Span. Zeitschrift für Erziehungswissenschaft, 14, 67–86.

Weisser, A. (2005). 18. Juli 1961 – Entscheidung zur Gründung der Ruhr-Universität Bochum. Web page, Internet-Portal Westfälische Geschichte.

Werblow, A., Felder, S., and Zweifel, P. (2007). Population Ageing and Health Care Expenditure: A School of 'Red Herrings'? Health economics, 16(10), 1109–1126.

Westphal, M. (2017). More Teachers, Smarter Students? Potential Side Effects of the German Educational Expansion. Ruhr Economic Papers 721, RWI Essen.

Westphal, M., Kamhöfer, D. A., and Schmitz, H. (2017). Marginal Labor Market Returns to Higher Education. Mimeo, Department of Economics, Paderborn University.

Wisselgren, M. (2005). Att föda barn – från privat till offentlig angelägenhet. Tech. rep., Umeå University.

Wissenschaftsrat (1960). Empfehlungen zum Ausbau der wissenschaftlichen Einrichtungen. Teil 1: Wissenschaftliche Hochschulen. Report, Wissenschaftsrat (German Council of Science and Humanities), Bonn.

Wissenschaftsrat (1966). Empfehlungen zur Neuordnung des Studiums an den wissenschaftlichen Hochschulen. Report, Wissenschaftsrat (German Council of Science and Humanities), Bonn.

Wissenschaftsrat (1970). Empfehlungen zur Struktur und zum Ausbau des Bildungswesens im Hochschulbereich nach 1970. Report, Wissenschaftsrat (German Council of Science and Humanities), Bonn.

Wolf, D. A., and Soldo, B. J. (1994). Married Women's Allocation of Time to Employment and Care of Elderly Parents. The Journal of Human Resources, 29(4), 1259–1276.

Wooldridge, J. M. (2016). Should Instrumental Variables be Used as Matching Variables? Research in Economics, 70(2), 232–237.

Zimmerman, S. D. (2014). The Returns to College Admission for Academically Marginal Students. Journal of Labor Economics, 32(4), 711–754.

Zimmermann, S., Artelt, C., and Weinert, S. (2014). The Assessment of Reading Speed in Adults and First-Year Students. Tech. rep., Leibniz Institute for Educational Trajectories (LIfBi) – National Educational Panel Study.

Zweifel, P., Felder, S., and Meiers, M. (1999). Ageing of Population and Health Care Expenditure: A Red Herring? Health economics, 8(6), 485–496.

List of Tables

1.1 Descriptive statistics. 28
1.2 Occupation of the household head according to the HISCO classification. 28
1.3 Regression results: second generation mortality. 34
1.4 Regression results: second generation mortality (50-70). Robustness and effect heterogeneity. 37
1.5 Regression results: mortality by cause of death. Robustness and effect heterogeneity. 38
1.6 Regression results: earnings in 1970. 42
1.7 Regression results: employment in 1970. 43
1.8 Regression results: years of schooling. 45
1.9 Regression results: employment in public services in 1970. 46
1.10 Assessing the potential selection effects: selection into fertility and in utero. 48
1.11 Regression of the infant mortality rate (IMR) on potential drivers (on regional level). 50
1.12 Assessing the the underlying channel behind the second generation mortality: infant mortality rate (IMR) in utero versus in early life. . . 51
1.13 Regression results, first generation mortality between 50 and 70. . . . 52
1.14 Cox proportional hazard regressions for mothers. 53
1.15 Descriptive statistics of the IMR driver analysis 56
1.16 Characteristic comparison: subsample dying before 1970 and subsample surviving to 1970. 56
1.17 Attrition in Income: Regression of log labor income in 1970 with IPW. 57

2.1 Stratified sample 65
2.2 Sample size 67
2.3 Care duration 68
2.4 Descriptive statistics according to treatment and matching status . . . 71
2.5 Table of results 79
2.6 SF-12v2 questionnaire in the SOEP 80

2.7 82
2.8 Distribution of p_{ij} across control variables in the sample 89
2.9 Distribution of p_{ij} across control variables in the sample 90

3.1 Numbers of observation 96
3.2 Sample means of outcome variables by care status 98
3.3 Summary of results 118
3.4 Variable description 123
3.5 Matching results corresponding to Figure 3.4 126
3.6 Variable selection, exemplary case: full-time work, static version . . . 131
3.7 Parameters for calibration of the sensitivity analysis 132
3.8 Back-of-the-envelope calculation of fiscal effects 134

4.1 Comparison of regions with and without college openings before college opens using administrative data 147
4.2 Descriptive statistics dependent variables 153
4.3 Descriptive statistics of instruments and background information . . 157
4.4 Regression results for OLS and first stage estimations 158
4.5 Estimated treatment parameters for main results 164
4.6 Potential mechanisms for cognitive skills 167
4.7 Potential mechanisms for health 168
4.8 Control variables and means by college degree 173
4.9 Full results for logit estimation of the selection equation (mean marginal effects) 177
4.10 Full results for 2SLS second-stage estimations 178
4.11 First-stage estimations when using different kernel bandwidths 182

5.1 Balancing test of regions with and without a college opening in the time under review using administrative data 195
5.2 Descriptive statistics of dependent variables 197
5.3 First stage and some characteristics of complying mothers 202
5.4 Baseline regression results 204
5.5 Post-college career outcomes as potentially mediating forces 211
5.6 Post-college family characteristics as potentially mediating forces . . 212
5.7 Control variables and means by university degree 216
5.8 Descriptive statistics of instruments and background information . . 217
5.9 Baseline fertility rates and college effects by age 218

6.1 Setup of the difference-in-differences approach 226
6.2 Descriptive statistics 234
6.3 Fixed effects results for math and reading competence 235
6.4 Main results – impact of the educational expansion on students' test scores 237
6.5 Robustness checks – placebo regression and predicting parental characteristics 238
6.6 Driving force behind effect 240
6.7 Tansformed results 247
6.8 Teacher mobility between federal states 247
6.9 Example structure of the data for the triple differences estimation . . 248
6.10 Number of pupil and number of students used in this analysis and dropping reasons 248
6.11 Descriptives for aspects of teacher's job choice 249
6.12 The association between the degree of the relative degree of the educational expansion and the Abitur grade for different samples 249
6.13 Potential impact of teacher non-response on the main effect 250
6.14 Sensitivity of the results due to selective student non-response and test participation 253
6.15 Potential impact of student non-response on the main effect 253

List of Figures

1 Descriptives of the demographic change in Germany (including GDR) 6

2 Descriptives of the educational expansion in Germany 7

1.1 Maps of first- and second-generation birth parishes. 26

1.2 Average infant mortality rate (IMR) by maternal year of birth. 30

1.3 Crisis indicator: average negative log deviation from 1930 tax revenues. 31

1.4 Second generation mortality hazards in the first year of life. 33

1.5 Second generation mortality hazards over the entire life cycle. 34

1.6 Cohort trends in mortality between 50 and 70, by sex and socioeconomic status (SES). 40

1.7 Distribution of the maternal disease environment by grandparental socioeconomic status (SES). 48

1.8 Distribution of the maternal disease environment by parental and grandparental socioeconomic status (SES). 49

1.9 Share of mothers with a bad health environment for selected parishes. 55

2.1 Basic time structure 64

2.2 Group assignment rules 65

2.3 Baseline results MCS and PCS 72

2.4 Alternative definitions of treatment and control groups (MCS only) . . 73

2.5 Alternative definitions of treatment and control groups (MCS only) . . 74

2.6 Results of the sensitivity analysis (MCS) 77

2.7 Results for an alternative specification of the propensity score vs. basline specification (2hrs MCS only) 83

2.8 Subgroup analysis for unmarried women with at least one parent alive (2 hours, MCS) 84

2.9 Alternative definitions of treatment and control groups (4 hours of care provision, MCS) 84

3.1 Time structure of the data 95

3.2 The distribution of care spells 97

3.3 Static design . . . 99
3.4 Labor market effects of informal caregiving for females – Static version 104
3.5 Pre-treatment trends for full-time employment . . . 106
3.6 Parameters for calibration of the sensitivity analysis . . . 109
3.7 Sensitivity analysis for full-time employment . . . 110
3.8 Dynamic design . . . 112
3.9 Labor market effects of informal caregiving for females – Dynamic version . . . 115
3.10 Results of the static version – Variations in treatment definition . . . 116
3.11 Matching quality for full-time work . . . 120
3.12 Static version, Kernel matching vs. IPW estimators . . . 120
3.13 Difference between baseline results and same estimation with restriction to never carers in the control group . . . 121
3.14 Results for females younger than 55 in $t = 1$ . . . 121
3.15 Double versus single post-lasso . . . 122
3.16 Exclusion of individuals with potential health shock or death of a parent between 0 and 1 . . . 127
3.17 Including labor market history and expectations . . . 129
3.18 Impact of assumed adverse measurement error, full-time work . . . 129
3.19 Results of the static version – Younger vs. older than median age . . . 130
3.20 Distribution of imputed care episodes . . . 135
3.21 Common support . . . 136

4.1 Average distance to the closest college over time for districts with a college opening . . . 145
4.2 Number of colleges and students over the time in selected states . . . 146
4.3 Distribution of propensity scores . . . 159
4.4 Marginal Treatment Effects for cognitive abilities and health . . . 161
4.5 Treatment parameter weights conditional on the propensity score . . . 164
4.6 Spatial variation of colleges across districts and over time . . . 170
4.7 Distribution of dependent variables by college graduation . . . 171
4.8 Sensitivity in Marginal Treatment Effects when using only the sum of the kernel weighted college distances . . . 172
4.9 Sensitivity in Marginal Treatment Effects when using different kernel bandwidths . . . 181

5.1 Trends in fertility and college enrollment by birth cohort in Germany 189
5.2 Mean age at first marriage and college enrollment by year in Germany 190

5.3 Colleges and students over time and by gender 192
5.4 Relative change in the share of students in counties within 100km of college opening between 1962 to 1969 201
5.5 Timing of births 208
5.6 Spatial variation of colleges across districts and over time 214
5.7 Trends in academic secondary school and college education for females 215
5.8 Trends in colleges and female students across federal states 215

6.1 Impact of the German educational expansion 223
6.2 Illustration of the fixed effects setup 224
6.3 Relative changes in the stock of teachers by non-urban federal states over time 225
6.4 Possible impact of the educational expansion on the job market for teachers 229
6.5 Number of teachers by year and subject 232
6.6 Cluster analysis of aspects teachers' job choice 242
6.7 The educational expansion and teachers' characteristics 242
6.8 Further characteristics of the educational expansion in Germany . . . 251
6.9 Number of Gymnasium teacher over time 251
6.10 Sensitivity of the effect with respect to the assignment year 252