RUPRECHT-KARLS-
UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Essays in Empirical Labor Economics

Evidence on Health, Education and Migration

Inauguraldissertation zur Erlangung des Doktorgrades
des Fachbereichs Wirtschaftswissenschaften der
Ruprecht-Karls-Universität Heidelberg

vorgelegt von
Dipl.-Volksw. Anna Lena Busse

Heidelberg 2019

BETREUERIN DER DISSERTATION

Prof. Christina Gathmann, Ph.D.

Lehrstuhl für Arbeitsmarkt und Neue Politische Ökonomie
Alfred-Weber-Institut für Wirtschaftswissenschaften
Universität Heidelberg

To my best friend and husband, Jan.

Acknowledgments

I would like to thank the various people that have guided and supported me during the time of my research.
First and foremost, this dissertation would not have been possible without my supervisor, Prof. Christina Gathmann, Ph.D. I especially want to thank Christina Gathmann for her advice, example and a very good balance between freedom and direction.
Secondly, I would like to express my gratitude to Prof. Dr. Stefan Klonner for serving as the second reviewer to this dissertation and would like to take this opportunity to thank him for providing a great example in creating a challenging yet constructive and inclusive working environment. I also want to thank him for two extremely helpful pieces of advice I often remember in moments of doubt in research and life in general.
I am also grateful to Prof. Dr. Maarten Lindeboom and many members of the School of Business and Economics at the Free University of Amsterdam who also hosted and supported me during this dissertation.
I thank my co-author Christina Vonnahme for her diligent work and great collaboration on what became the third chapter of my dissertation and the second of hers.
During my time in Heidelberg, I was lucky to be well accompanied and supported by my colleagues at the Chair of Labor Economics and New Political Economy and the Faculty of Economics and Social Sciences. Many of whom have contributed to my professional success and personal development in very important ways. I am particularly grateful for those who entered my life as colleagues and continue to be part of it as friends.
Finally, and on a personal note, I would like to take this opportunity to acknowledge the contribution of the people who have been part of my life long before the work on this thesis began - my beloved parents, sister, family and friends. I consider myself most fortunate to be able to be with so many wonderful people.

Contents

List of Tables

List of Figures

Acronyms

ATE Average Treatment Effect

ATT Average Treatment Effect on the Treated

ATUT Average Treatment Effect on the Untreated

AuslG Ausländergesetz (Alien Act)

CDU Christian Democratic Union

CMS Compulsory Military Service

DRG Diagnosis Related Group

FDP Free Democratic Party

FTE Full-Time Equivalences

GDP Gross Domestic Product

KHEntgG Krankenhausentgeltgesetz (Hospital Remuneration Act)

ICD International Classification of Diseases

ITT Intention to Treat Effect

KHG Gesetz zur wirtschaftlichen Sicherung der Krankenhäuser und zur Regelung der Krankenhauspflegesätze (Hospital Financing Act)

LATE Local Average Treatment Effect

LIV Local Instrumental Variable

MPRTE Marginal Policy Relevant Treatment Effects

MTE Marginal Treatment Effect

NEPS National Educational Panel Study

OECD Organisation for Economic Co-operation and Development

PISA Programme for International Student Assessment

SDQ Strengths and Difficulties Questionnaire

SOEP Socio-Economic Panel

SPD Social Democratic Party

StAG Staatsangehörigkeitsgesetz (Citizenship Act)

VABS Vineland Adaptive Behavior Scale

Introduction

"The master-economist must possess a rare combination of gifts. ... He must be mathematician, historian, statesman, philosopher in some degree. He must understand symbols and speak in words. He must contemplate the particular, in terms of the general, and touch abstract and concrete in the same flight of thought. He must study the present in the light of the past for the purposes of the future. No part of man's nature or his institutions must be entirely outside his regard. He must be purposeful and disinterested in a simultaneous mood, as aloof and incorruptible as an artist, yet sometimes as near to earth as a politician."

John Maynard Keynes

Research in economics is closely related to public policy. This applies particularly to labor economics where researchers often aim to shed light on the causal effect of either a policy intervention (for example the introduction of a minimum wage or the abolishment of a tuition) or an individual choice variable (e.g. child care utilization) on labor market outcomes such as labor force participation or income (Van Der Klaauw 2014). Questions like "Does free child care facilitate maternal labor force participation?", "Does employing people performing community service in hospitals support patient health?" or "Who benefits from a liberal citizenship policy?" are equally important to economists and politicians.

To illuminate such questions it is important to provide reliable econometric estimates on the relationship between causes and effects. This, however, can be quite challenging. Considering the potential outcomes model developed by Rubin (1974), based on earlier work by Neyman (Neyman and Iwaszkiewicz 1935), helps to understand why. The model illustrates that each individual can be observed in two states: the state where it did not obtain treatment and the state where it did. One can think of *treatment* as a medical treatment or a change of rules and regulations - a policy change. Rubin (1974) calls

these states *potential outcomes*. The difference between these two potential outcomes is the causal effect for an individual of participating in the treatment. Since each person is either treated or untreated only one of the two potential outcomes can be observed. Thus, it is simply impossible to directly observe the treatment effect for an individual. Instead, the treatment effect is unobservable which the economist calls an *unobserved random variable*. Holland (1986), one of Rubin's students, famously described this fact as the *fundamental problem of causal inference*.

The gold standard for the identification of a causal effect of an intervention is the observation of treated and untreated control units in a randomized controlled experiment (Athey and Imbens 2017). Such experiments provide data where by randomization the units that are exposed to an intervention, i.e. treated, in expectation are the same as those who are not. The difference between the average outcome of the treated and untreated can be interpreted as an unbiased estimate of the average causal effect (Average Treatment Effect (ATE)). Thus, these experiments overcome the fundamental problem of causal inference by approximating the individual treatment effect by an average treatment effect. Initially conducted in medical and psychological science, economic research has discovered the value of randomized controlled lab and field experiments decades ago (Deaton and Cartwright 2018). Yet, there are many settings where experimentally testing policy interventions in randomized experiments is too costly or politically not feasible. Think, for instance, of randomly allocating citizenship to migrants to learn about the effects of naturalization on educational success or forcing hospitals to hire or lay off medical staff to study how labor and capital interact in treating patients. Still, policy decisions on these and similar questions have to be made and reliable empirical evidence is needed to inform these decisions. Economists, in particular labor economists, have been very active in developing and adapting microeconometric methods to provide evidence in settings where controlled experiments are not feasible (Van Der Klaauw 2014).

This thesis consists of three essays drawing on these concepts. It provides empirical evidence from actual policy changes in rather distinct areas of labor economics: health, education and migration. Each essay analyses an actual policy change and exploits a different identification strategy to derive empirical estimates on economic questions relevant to policy makers.

The first essay *"Low-Skilled Labor in Hospital Production: Evidence from the Suspension of Compulsory Military Service in Germany"* analyses the relationship between low-skilled labor, capital input and hospital production. In particular, it discusses questions like "How do low-skilled employees contribute to health care production in the highly professionalized setting of German hospitals?" and "How does low-skilled labor input

interact with capital input as, for instance, medical equipment and drugs?" and "How does a different ratio between labor and capital input alter the productivity of hospitals for example in terms of patient's health (numbers of patients treated successfully, death rates etc.)?".

The second essay *"Free Universal Daycare: Effects on Children and Maternal Labor Supply"* analyses the effects of free universal daycare on childcare attendance, maternal labor supply and child development. Specifically, it discusses questions like "How does the introduction of free public daycare affect childcare arrangements?", "Does offering free daycare facilitate maternal labor supply?" and "What is the effect of access to free daycare on children?".

Finally, the third essay *"Marginal Returns to Citizenship and Skill Development"* provides evidence on whether the option to naturalize improves educational outcomes of migrant's children. It sheds light on questions like "Who takes the opportunity to naturalize when citizenship policy is liberalized?", "Who benefits from citizenship in terms of educational outcomes?"and "Under which circumstances could a reform of immigration policy carry additional net benefits for children?".

Gaining insights in these questions is relevant to economists and policy makers alike. Let us consider why for each essay in more detail:

Consider, for instance, the first essay *"Low-Skilled Labor in Hospital Production: Evidence from the Suspension of Compulsory Military Service in Germany"*. Labor is not only a very crucial input factor in health care production, labor of health care professionals is also an important economic factor. Health worker's wages account for more than half of health care spending, whereas health care sector employment accounts for 10% of the total workforce in OCED-countries (Organisation for Economic Co-operation and Development) (OECD 2016). Considering the growing demand for health and care services, it is of increasing importance to assess how different types of labor function and interact, for one thing, with each other and, for another thing, with capital input in health care production. Yet, despite the importance and topicality of this issue, data limitations and endogeneity concerns have largely prevented the existing literature from providing clear cut evidence on the causal relationship between hospital staffing and production (Friedrich and Hackmann 2017).

A similar line of argument applies to the second essay *"Free Universal Daycare: Effects on Children and Maternal Labor Supply"*. Many governments have expanded their social policies in the area of early childcare and education. While countries like France, Sweden, Norway or Denmark have long offered universal access to public childcare, others like Germany, Spain, Canada or the US have expanded public daycare and pre-K programs much more recently - beginning in the 1990s. Proponents of such policies argue that money

invested in early childhood education is well spend as it would simultaneously boost the human capital development of preschool children and encourage female labor supply. Yet, the empirical evidence on the link between daycare availability and maternal labor supply is still mixed.

The third essay discusses *"Marginal Returns to Citizenship and Skill Development"*: Over recent decades, many developed countries have accumulated sizable immigrant populations. In 2017, the foreign-born made up 15.5 percent of the population in Germany, almost 13 percent in France and even close to 30 percent in Switzerland (OECD 2018a). At the same time, immigrants often seem to perform poorly compared to natives and sometimes are not able to catch up over generations. Previous evidence has shown that a liberal citizenship policy improves the labor market position, especially of immigrant women (Gathmann and Keller 2018). Yet, it remains unclear if positive effects on the first generation of immigrants carry over to immigrants' children. This is relevant to ensure equal opportunities for migrants and natives alike but is also of interest to policy makers because a disadvantaged economic position reduces the fiscal benefits of immigration (Dustmann and Glitz 2011). Systematic differences might threaten the social cohesion of host countries producing hostility among natives as well as immigrants.

While each of the essays contributes to a different literature in the field of labor economics all three of them rely on observational data, i.e. data from policies that were not implemented randomly (Athey and Imbens 2017). Potential data sources containing the necessary information are, for instance, surveys (Socio-Economic Panel (SOEP), National Educational Panel Study (NEPS)), administrative records (Hospital and Patient Statistics) or (digital) protocols (consumption data, scraped data). The identification of causal effects from observational data, however, can be quite challenging.[1] Consider, for example, the introduction of minimum wages, as discussed many times in economic literature, very recently also in a paper by Athey and Imbens (2017): A researcher interested in the causal effect of minimum wages on employment might observe the introduction of high minimum wages on a state level in some states while other states do not implement any or only low minimum wages. It is plausible that states with higher costs of living and less price-sensitive consumers are those that implement higher minimum wages. At the same time, employers in these states might also be more able to pass on higher production costs to consumers without losing too much business. This does not need to be true in states where living costs are low and consumers are more price-sensitive. The latter states might choose a lower level of minimum wages. If a researcher interested in the effect of a higher minimum wage on employment followed the approach of randomized controlled studies

[1] Refer to Deaton and Cartwright (2018) for some reflective thoughts on causal inference based on randomized controlled experiments.

and naively compared the average employment level of states with a high minimum wage to that of states with a low minimum wage he or she would *not* obtain a credible estimate of the causal effect of an increase in minimum wages.

Observational data almost never provide a control group that, in expectation, is the same as the treatment group offhand. In the vast majority of cases the policy is not randomly implemented, data on treatment as well as potential control units is possibly confounded by simultaneous events and the independent variable is not manipulated by the researcher. The key to success with observational data is to search for *exogenous variation* in treatment assignment to obtain a *natural experiment* or *quasi-experiment*, i.e. variation in treatment participation that is not related in any way to the outcome of interest (Van Der Klaauw 2014). A starting point for such a search could, for instance, be an age based cut-off rule, regional or temporal variation in policy implementation or variation in treatment intensity due to pre-determined characteristics. Often exogenous variation stems from institutional rules causing almost identical individuals to be exposed to different treatment schemes (Van Der Klaauw 2014).

The first essay of this dissertation, for instance, uses the suspension of military drafting in Germany to advance our understanding of labor in hospital production. How can the suspension of drafting help shedding light on this? Germany followed the international trend towards a professional army when it indefinitely suspended the compulsory military service in 2011. As a consequence, the number of draftees decreased dramatically over the course of a year. This also led to a sudden drop in hospital staffing because by law (*Zivildienstgesetz*) draftees were allowed to serve in a *"position for the common good"*, i.e. perform some type of community service instead of joining the armed forces ("conscientious objection"). Opting for community service was very common such that, at the time of the suspension, 6% of the care staff in general hospitals were young men performing community service. Yet, not all hospitals employed people performing community service before the policy change. This enables me to divide the population of German hospitals into a treatment and a control group.

The second essay of this thesis exploits regional variation as well as age and temporal variation in policy implementation from the adoption of free daycare policies in West Germany between 2007 and 2016. Regional variation stems from the fact that nine out of eleven states adopted a free daycare policy. The timing of adoption within states generates additional variation by birth cohorts of children. Finally, the policies cover different age groups. While all reform states introduced free daycare for the last year prior to school entry ("kindergarten"), some states introduced more comprehensive reforms covering children aged between 2 and 5 ("pre-K"). Consequentially, children who were not eligible for free child care because of their age or place of residence can serve as a control

group here.

Finally, the third essay relies on variation in eligibility for citizenship across migrants' arrival year in Germany and their birth cohort. The variation is induced by two reforms in citizenship law which together defined four routes to citizenship for children of immigrants: birthright citizenship, an associated transitional rule, eligibility through parents and individual eligibility. The first reform, taking place in 1991, defined age-dependent residency requirements for naturalization. Consequently, younger and older immigrants arriving in the same year potentially faced different waiting periods until they were allowed to apply for citizenship. Additional variation is obtained from the second reform, implemented in 2000, that allowed all immigrants to apply for citizenship after 8 years regardless of their age and on top of that introduced citizenship by birth. The estimation approach then implements a marginal treatment effects framework using the various eligibility indicators as instruments.

Identifying feasible exogenous variation is key but only the first step. To establish causal effects exogenous variation has to be paired with the appropriate econometric approach which jointly is often described as an *identification strategy*, i.e. a way to use observational data to approximate a randomized controlled experiment. To be able to provide credible estimates of causal effects from observational data labor economists have been very active in developing and refining micro econometric methods and identification strategies in the last fifty years. By this time, the policy evaluation toolbox contains several non-experimental and quasi-experimental identification strategies. Around these methods and strategies a sizeable literature has evolved. The same is true for their applications. Van Der Klaauw (2014) provides a brief history of the methods developed in empirical labor economics and discusses several important applications for each of them, whereas Athey and Imbens (2017), Imbens and Wooldridge (2009), Heckman and Vytlacil (2007a), Heckman and Vytlacil (2007b) among others, discuss the recent developments in causality and policy evaluation, at times even in textbook extension (Imbens and Rubin 2015). The choice of the appropriate approach eventually depends on the treatment assignment mechanism and data availability (Van Der Klaauw 2014).

The first and second essay in this thesis rely on variants of a difference-in-difference approach, whereas the third essay exploits a marginal treatment effect framework to identify the effect of the policy. Thus, all three essays presented in this thesis build on this by now rather mature literature. To see how the papers arrive at their results and how they relate to the previous work in their particular field of literature let us consider each of the essays in more detail:

Low-Skilled Labor in Hospital Production:
Evidence from the Suspension of Compulsory Military Service in Germany

The analysis of natural experiments demands detailed information regarding cause and context of the exogenous variation and a sufficient number of observations at the margin of the natural experiment (Van Der Klaauw 2014). Oftentimes, surveys do not fulfill these requirements. Thus, facilitated by the increasing digitization of administrative records, using administrative data became increasingly popular in the last years. Chapter 1 of this dissertation follows this trend by using administrative data from the German Hospital and Patient Statistics, a data set established to monitor and plan hospitals by governmental authorities. This data set provides detailed information on the universe of general hospitals of the country and the vast majority of their patients; for some outcomes even with daily precision.

This essay provides results on three different sets of outcomes. Firstly, I show that hospitals were able to maintain their output levels in terms of quantity as well as in terms of measurable quality, when people performing community service dropped out of their workforce. This applies to a wide set of objective outcomes, such as numbers of in- and out-patients treated and overall care days performed. Considering patient outcomes shows that neither patients' duration of stay nor their likelihood of dying in the hospital or being referred to another hospital was affected on the hospital level. However, when individual patients are considered, there is some evidence for an adverse effect on durations of stay. Although statistically significant it is not sizable enough to level the overall time trend of decreasing durations of stay. Secondly, medical institutions did not alter their contracted labor input in measurable ways. The loss in people performing community service was neither compensated by training more apprentices nor by hiring more nurses or any other staff group related to patient treatment. Yet, thirdly, evidence on hospital expenditure shows that staffing costs increased, either due to hours worked overtime or higher salaries. I also show that expenditure on material costs (including medical consumables like drugs, bandages, instruments, therapeutic appliances, blood and plasma) went up as a consequence of the drop out of low-skilled labor, indicating that hospitals switched to a more capital intensive production. This is of interest because a more capital intensive production, for instance through increased drug administration, has been associated with detrimental effects on patients' health (e.g. Cawley et al. 2006: on nursing homes).

Using the suspension of drafting as a natural experiment is the main contribution of this paper and has a number of new and interesting features when studying labor input in hospital production. Firstly, it did not originate in the health system. Instead, it was implemented to decrease military expenditures[2]. Earlier empirical evidence is mostly based

[2]Another reason was the aim to re-establish fairness in drafting among young men. Although in principle

on (seemingly) exogenous variation caused by fluctuations in patient loads (Evans and Kim 2006), minimum staffing regulations (Cook et al. 2012; Tong 2011) or on strikes in the health system (Gruber and Kleiner 2012; Cunningham et al. 2008). The disadvantage of these variations, caused within or as a reaction to the situation in the health system, is that they can be foreseen and accounted for, potentially causing endogeneity problems that downward bias the point estimates. During physician strikes in Israel, for instance, doctors refused to treat patients in hospitals but established separate aid stations for treatment outside (compare Cunningham et al. 2008). One other notable exception to this literature is provided by Friedrich and Hackmann (2017) who study the effects of a parental leave program offered to Danish parents on health care delivery. Secondly, the policy change was implemented rather quickly and led to a sharp drop in low-skilled labor. It was debated in fall 2010 and already voted for in parliament in December of the same year. The last men were drafted in the following June such that by the end of 2011 there were no more men performing military or community service compulsorily. As a consequence, the number of people performing community service dropped from around 80,000 in the previous years to 0 by the end of 2011, leaving hospitals little time to adapt to the policy change beforehand and thus providing profound variation to study. Thirdly, this indefinite suspension of drafting enables me to study a much longer time horizon than most other studies, as strikes or seasonal variations are typically short-lived.

Free Universal Daycare:
Effects on Children and Maternal Labor Supply

In this essay, we exploit the staggered introduction of free universal daycare in several German states to track how families with pre-school children between the age of 2 and 6 respond to and benefit from the policies. Our analysis yields five main findings. First, the free daycare policy only affects the youngest children aged between 2 and 3. Access to a free daycare slot raises daycare attendance in that age group by 8.4 percentage points or 17% relative to the pre-policy period.

Second, daycare attendance for older children (aged 3 and above) does not respond to the free daycare policy. In particular, we find no effect for the most common policy that adopted a free year of daycare prior to school entry ('kindergarten'). The main reason is that daycare attendance for children aged 3 and above has been high (94.4 percent) even before the policies were introduced.

Third, we observe substantial positive employment effects among mothers with 2-3 year-

all able-bodied men could be drafted, eventually, a decreasing share of men were actually committed. In April, 21, 2004 the administrative court in Cologne for the first time raised doubts about fairness in drafting. This case, however, was not negotiated through all official channels before the suspension of drafting. For more details on this topic refer to Fleischhauer (2007).

old children after the policy is adopted. Labor force participation increases by 7.7 percentage points or 17% relative to the pre-reform period. Mothers with older children (aged between 4 and 6) in turn increase their labor supply at the intensive margin – working more full-time and increase working hours by almost two hours per week (or 11%). Overall, these results imply that the additional income saved from the free daycare slot is not used to reduce female labor supply.

Fourth, the free daycare policy has few persistent effects on child development. For the youngest children, we observe no overall effect, though a negative effect on skills in daily activities (i.e. whether the child can use a spoon on its own, for instance) and a positive effect on social skills. For children aged 5 to 6, we find no overall effect and also no effect on sub-categories measuring behavioral problems. Hence, the policy seems to do no persistent harm, but also creates few benefits in terms of cognitive or non-cognitive skills for the average preschool child.

Finally, we document substantial heterogeneity in the treatment effect: poor and low-skilled households respond more to the policy than the average family. Children from low-skilled and poor households are more likely to attend public daycare and less likely to be cared exclusively at home. Poor children in particular benefit from free daycare, which boosts their cognitive and non-cognitive skills. In contrast, we find no effect on female labor supply suggesting that either the returns to work (more) or labor supply elasticities more broadly are small for low-income and low-skilled mothers.

With this analysis we make four important contributions to the literature on childcare and early childhood education. Most importantly, we can compare whether the margins of adjustment to and benefits of a free daycare policy differ for 2-year-olds and 5-year-olds, for instance. Since attendance rates in many countries have traditionally been much higher for older pre-school children, providing daycare free of charge is likely to generate stronger behavioral responses among families with very young children for which attendance rates are still low.

Second, we can assess the impact of free daycare on a range of family choices, including childcare arrangements and labor supply, as well as short-run child development. Analyzing the full range of family responses to daycare policies is crucial for interpreting the estimated effects on child development. Children are less likely to benefit from a free daycare policy, for instance, if parents switch from high-quality parental care to low-quality public care; they are more likely to benefit if families switch from low-quality informal (or parental) care to high-quality daycare instead.

Because the policies we study are universal, our third contribution is to shed light on who responds to the policy; and whether the benefits accrue to the average child or are concentrated among children from disadvantaged families, for instance. Whether public

daycare is able to level the playing field between poor and more affluent families is of central interest to policy-makers concerned about equality of opportunity in early childhood education.

Finally, the policy we analyze offered childcare subsidies rather than merely expanding daycare supply. Providing additional daycare slots expands the choice set of parents, which generates substitution effects out of care at home or informal care into formal childcare. Free daycare, in turn, is equivalent to a price decline of public daycare relative to other childcare options and, since it is uncompensated, there will be both income and substitution effects on childcare and labor supply choices. Hence, children might still benefit from the policy even though parents do not adjust their labor supply behavior, for instance.

Marginal Returns to Citizenship and Skill Development

Finally, the third paper estimates the causal effects of citizenship and explores its heterogeneity for the skill development of children and young adults along both observable and unobservable characteristics. Our main data source is the National Educational Panel Study (NEPS, Blossfeld et al. 2011), which is particularly suited for our analysis as the panel collects detailed information on children's education and skill development. A unique feature of the data for our purpose is that the NEPS study administers standardized competence tests to all children. Such detailed standardized test results are rarely available in Germany and certainly not for large samples of school-aged children. We focus in our analysis on young immigrants as immigrant-native gaps in language, math and science tests are sizable and persistent as has been demonstrated, for instance, for 15-year-olds (Dustmann and Glitz 2011; OECD 2018b). Similarly, the ethnic gap in achievement test scores and other skills increases substantially during childhood (Fryer Jr. and Levitt 2004; Heckman et al. 2006; Cunha and Heckman 2007). Given the strong link between student competencies and outcomes later in life, these gaps are likely to have long-term effects on the occupational careers and labor market success of children from immigrant families, reducing upward mobility and cementing unequal opportunities (Dustmann and Glitz 2011).

Indeed, we find substantial heterogeneity in the returns to citizenship along both observable and unobservable characteristics. Immigrant children born in Germany and girls in particular benefit substantially from German citizenship in terms of language skills as measured by standardized test scores but also their school grade in German. At the same time, German-born children of immigrants are more likely to naturalize pointing to a positive selection on gains. Immigrant girls in turn, are just slightly more likely to obtain citizenship than immigrant boys. The positive selection in gains is reflected in unobserved heterogeneity as well: returns to citizenship with respect to language skills are declining

with increasing resistance to treatment. Hence, children whose unobserved characteristics make them most likely to naturalize benefit the most, while children whose unobservable characteristics make them least likely to pick up German citizenship have zero returns to host country citizenship. We further show that improvements in language skills also help immigrant children to improve their school performance: children are much less likely to repeat a grade in school, for instance.

Our analysis makes several important contributions to the literature. Existing studies on the consequences of citizenship have estimated at most intention-to-treat effects of birthright citizenship on early childcare attendance, school entry and school track (Felfe et al. 2019). Our analysis estimates causal effects (Local Average Treatment Effect (LATE)), of citizenship on skill development throughout primary and secondary school. Even more importantly, we explore for the first time the heterogeneity of returns to citizenship across observable and unobservable characteristics by estimating Marginal Treatment Effects. Furthermore, our data cover all states in Germany and allows to control for detailed parental characteristics like years since migration, which is typically not available in administrative datasets.

A second line of research analyzes how birthright citizenship has affected parental integration efforts (Avitabile et al. 2013), fertility behavior (Avitabile et al. 2014) and return migration (Sajons 2016) showing that parents increase their efforts to use the local language and reduce the total number of children in favor of more investments in the 'quality' of children.

Another line of research investigates the consequences of access to citizenship on the labor market performance and marriage behavior of adult immigrants using panel data approaches (Bratsberg et al. 2002; Steinhardt 2012) or intention-to-treat effects (Gathmann and Keller 2018; Gathmann et al. 2019). Unlike all previous evidence, we focus on how access to citizenship through parents, individual eligibility or birthplace affects children's skill development, which is a crucial prerequisite for later success in the labor market. In addition, we are able to identify causal effects and explore for the first time their distribution along observable and unobservable dimensions.

Our study also contributes to the growing literature that estimates marginal treatment effects. Most studies focus on monetary returns to a college education (see e.g. Carneiro et al. 2011; Kaufmann 2014; Nybom 2017; Kamhöfer et al. 2018), to secondary education (e.g Carneiro et al. 2011: in Indonesia) or to early childhood education (Cornelissen et al. 2018; Felfe and Lalive 2018; Kline and Walters 2016).

CHAPTER 1

Low-Skilled Labor in Hospital Production: Evidence from the Suspension of Compulsory Military Service in Germany

1.1 Introduction

Labor is not only a very crucial input factor in health care production, labor of health care professionals is also an important economic factor. Health worker's wages account for more than half of health care spending, whereas health care sector employment accounts for 10 percent of the total workforce in OECD countries (OECD 2016). Considering the growing demand for health and care services, it is of increasing importance to assess how different types of labor input function and interact, for one thing, with each other and, for another thing, with capital input in health care production. Yet, despite the importance and topicality of this issue, data limitations and endogeneity concerns have largely prevented the existing literature from providing clear cut evidence on the causal relationship between hospital staffing and production (Friedrich and Hackmann 2017).

This paper uses the suspension of military drafting in Germany to advance our understanding of labor in hospital production. In particular, it discusses questions like "How do low-skilled employees contribute to health care production in the highly professionalized setting of German hospitals?" and "How does the input of low-skilled labor interact with capital input as, for instance, medical equipment and drugs?" and "How does a different relation between labor and capital input alter the productivity of hospitals for example in terms of patient's health?".

How can the suspension of drafting help shedding light on these questions? Germany followed the international trend toward professional armies when it indefinitely suspended the Compulsory Military Service (CMS) in 2011. As a consequence, the number of draftees

decreased dramatically over the course of a year. This also led to a sudden drop in hospital staffing because by law (*Zivildienstgesetz*) draftees were allowed to serve in a "*position for the common good*", i.e. perform some type of community service instead of joining the armed forces (conscientious objection). Opting for community service was very common, such that, at the time of the suspension, 6 percent of the care staff in general hospitals were young men performing community service.

Using the suspension of drafting as a natural experiment has a number of interesting features when studying labor input in hospital production. Firstly, it did not originate in the health system. Instead, it was implemented to decrease military expenditures[1]. This paper is among the first to use exogenous variation affecting hospital staffing but not originating in the health system. Earlier empirical evidence is mostly based on (seemingly) exogenous variation caused by fluctuations in patient loads (Evans and Kim 2006), minimum staffing regulations (Cook et al. 2012; Tong 2011) or on strikes in the health system (Gruber and Kleiner 2012; Cunningham et al. 2008). The disadvantage of these variations, caused within or as a reaction to the situation in the health system, is that they can be foreseen and accounted for, potentially causing endogeneity problems that downward bias the point estimates. During physician strikes in Israel, for instance, doctors refused to treat patients in hospitals but established separate aid stations for treatment outside (compare Cunningham et al. 2008). One notable exception to this literature is provided by Friedrich and Hackmann (2017) who study the effects of a parental leave program offered to Danish parents on health care delivery.

Secondly, it was implemented rather quickly and led to a sharp drop. The policy change was debated in fall 2010 and already voted for in parliament in December of the same year. The last men were drafted in the following June such that by the end of 2011 there were no more men performing military or community service compulsorily. As a consequence, the number of people performing community service dropped from around 80,000 in the previous years to 0 by the end of 2011, leaving hospitals little time to adapted to the policy change beforehand.

Thirdly, this indefinite suspension of drafting enables us to study a much longer time horizon than most other studies, as strikes or seasonal variations are typically short-lived. Using administrative data on all general hospitals of the country and the vast majority of their patients (German Hospital and Patient Statistics), this paper provides results on three different sets of outcomes. Firstly, it shows that hospitals were able to maintain their

[1] Another reason was the aim to re-establish fairness in drafting among young men. Although in principle all able-bodied men could be drafted, eventually, a decreasing share of men were actually committed. In April, 21, 2004 the administrative court in Cologne for the first time raised doubts about fairness in drafting. This case, however, was not negotiated through all official channels before the suspension of drafting. For more details on this topic refer to Fleischhauer (2007).

output levels in terms of quantity as well as in terms of measurable quality, when people performing community service dropped out of their workforce. This applies to a wide set of objective outcomes, such as numbers of in- and out-patients treated and overall care days performed. Considering patient outcomes shows that neither patients' duration of stay nor their likelihood of dying in the hospital or being referred to another hospital was affected on the hospital level. However, when individual patients are considered, there is some evidence for an adverse effect on durations of stay. Although statistically significant, it is not sizable enough to level the overall time trend of decreasing durations of stay.
Secondly, medical institutions did not alter their contracted labor input in measurable ways. The loss in people performing community service was neither compensated by training more apprentices nor by hiring more nurses or any other staff group related to patient treatment.
Yet, thirdly, evidence on hospital expenditure shows that staffing costs increased, either due to hours worked overtime or higher salaries. It also can be shown that expenditure on material costs (including medical consumables like drugs, bandages, instruments, therapeutic appliances, blood and plasma) went up as a consequence of the drop out of low-skilled labor, indicating that hospitals switched to a more capital intensive production. This is of interest because a more capital intensive production can be associated with detrimental effects on patients' health (compare Cawley et al. 2006: on nursing homes).
This chapter proceeds as follows. The following section provides a brief overview of the related literature. Section 3 describes the institutional background. Section 4 provides the empirical analysis, discusses some microeconomic considerations, potential effects as well as details on the applied data and the empirical approach. Section 5 then outlines the results, before concluding remarks are presented in the final section 6.

1.2 Literature

By exploiting exogenous variation from the suspension of drafting, this analysis contributes to two main strands of literature: the literature on labor in hospital production and the economic literature on conscription.
The literature on labor in hospital production includes, but is not limited to, an analysis of strikes by nurses and doctors. Evidence on strikes of nurses suggests that the consequential labor input reduction leads to an increased duration of stay and higher in-hospital mortality (Gruber and Kleiner 2012). Previous research on strikes of doctors, however, reveals a puzzling pattern of a reduction of work effort in hospitals. Evidence from several strikes, lasting between 9 days and 17 weeks, indicates that when physicians go on strike mortality levels stay constant or even decrease. Cunningham et al. (2008)

provide a comprehensive review of this literature. This counter-intuitive pattern is only partly explained by the fact that scarce staff is re-assigned to compensate for the loss of labor input and the fact that the number of elective surgeries are reduced. As the longest strike only lasted 17 weeks these studies, however, cannot rule out that this finding is a short-run effect. As the suspension of drafting in this analysis is indefinite (and still in place today), this paper can complement these previous studies by providing evidence from a more permanent reduction in staffing.

Besides the literature considering changes in labor input due to strikes, a number of studies have looked at the extensive margin of labor supply by studying patient-to-nurse ratios. In summary, they suggest that increases in patient-to-nurse ratios are associated with increases in mortality and other adverse effects on in-patients. Seminal papers in this area have been provided by Aiken et al. (2002) and Needleman et al. (2002). These first studies, however, rely on cross-sectional variation. This paper, alongside other studies, adds to this early work by relying on exogenous variation. Among the first to provide evidence from exogenous variation are Evans and Kim (2006) who use changes in patient loads. These authors find that patients, who are admitted when patient loads are high, tend to have higher mortality but effects are estimated to be quite small and are not statistically significant in several of their specifications. Cook et al. (2012) provide evidence from legally mandated increases in nurse staffing in California and do not find a discernible effect on patient safety. The same variation is also exploited by Tong (2011), who finds that increased staffing reduces on-site patient mortality, and Lin (2014), who establishes a positive relationship between number of registered nurses and quality of care in nursing homes. She also establishes that the potential endogeneity problem in previous cross-sectional studies would downward bias estimates and thereby could be responsible for low effect sizes established by seminal papers. While there is some evidence on the consequences of a short-run reduction in high-skilled and well-trained labor in hospitals there is only little evidence on the effects of changes in labor input by untrained staff. Lin (2014) is among the few studies that also provide evidence on nurse aid staffing which is most comparable to people performing community service. Nurse aid in that study did not have a significant influence on care quality.

The vast majority of studies relies on variations caused within the health system. One notable exception is provided by Friedrich and Hackmann (2017) who analyze the effect of a 10 percent drop in labor input in the Danish health system caused by the introduction of a parental leave program. These authors identify an increase in 30-day readmission rates and a distortion of technology utilization in hospitals while mortality remains unaffected in this study. Studying the effects of the same policy change in nursing homes, however, reveals a 13 percent increase in mortality among patients aged 85 or older. These dif-

ferential effects point to interesting differences in returns to nursing across sub-sectors of the health system. While the work by Friedrich and Hackmann (2017) relies on variation predominately affecting female and licensed professionals, this study adds to the literature by providing evidence from a large scale reduction in the labor input of low-skilled males, with potentially very different effects.

There is also a somewhat smaller literature evaluating the intensive margin of labor input. Bartel et al. (2014) provide evidence on the productivity of nurses' human capital. They study how education and unit-specific human capital from experience affect patients and show that both components of human capital significantly improve patient outcomes. Additionally, they demonstrate that disruptions to the care team (for instance the departure of experienced nurses, the absorption of new hires, and the inclusion of temporary contracted nurses) negatively affect productivity by increasing patients' durations of stay. Aiken et al. (2003) show that an increase in the proportion of nurses holding a bachelor's degree is associated with a decrease in both the likelihood of surgical patients dying within 30 days of admission and the odds of failure to rescue. Doyle et al. (2010) study how patient outcomes vary between well and less well trained doctors and interestingly find no effect on health but uncover that treatment by less skilled doctors is more costly, thus implying that hospitals move to a production point with a higher capital to labor input ratio when operating with lower human capital. The results of this paper complement the results of Doyle et al. (2010) by indicating that not only changes in the intensive margin of high-skilled labor but also changes of the extensive margin of low-skilled labor input lead to significant adaptations in the capital and labor input structure of hospitals.

By providing evidence on the interplay of labor and capital in hospital production this paper also relates to the literature on factor substitution. Besides numerous papers studying factor substitution in general[2] there are some previous papers studying factor substitution in the health sector which are related to this paper in relevant ways. Most importantly, the papers by Acemoglu and Finkelstein (2008) and Cawley et al. (2006) who both provide evidence from the U.S.American health system. Acemoglu and Finkelstein (2008) provide evidence from hospitals showing that an increase in labor costs leads to lower labor input and induces a higher capital to labor ratio. Cawley et al. (2006) provide similar evidence from nursing homes. Specifically, they show that higher nursing home wages are associated with greater use of psychoactive drugs and lower care quality. This paper complements the already existing literature in two ways. By providing evidence on factor substitution induced by a drop in labor supply and by providing evidence from the German setting. To the best of my knowledge, there is just one previous working paper considering factor substitution in the German hospital sector. The work of Schmitz and

[2]Google Scholar provides more than 3.5 million results on "factor substitution".

Tauchmann (2012), however, is very different from this paper as the authors specifically focus on the establishment of ownership-specific heterogeneity in the technical elasticity of substitution between physicians and nurses in the wake of new regulations regarding the remuneration of hospitals implemented in the second half of the last decade. Their findings indicate that non-profit hospitals operate particularly physician-intensive.[3]
By exploiting exogenous variation from the suspension of mandatory military service this paper also contributes to the understanding of the economic effects of drafting and its substitute services. The volume of drafts (all able men in each cohort), the duration (between 6 and 24 months in most countries), and the fact that young men and, in some countries, women are drafted during a period in their lives that is usually used for human capital investment[4] (commonly starting at the age of legal majority) indicate, that drafting can have potentially far reaching consequences for economies as a whole as well as for recruits personally. Thus, the economic literature has long been considering the concept of conscription, beginning with general considerations of the costs and inefficiencies of drafting decades ago (Fisher 1969; Oi 1967). In particular, the consequences for the recruits themselves in terms of labor market outcomes (e.g. Bauer et al. 2012; Grenet et al. 2011; Imbens and Klaauw 1995), demand for education (e.g. Bauer et al. 2014; Card and Lemieux 2001) and health (Angrist et al. 2011) are well researched. There is also a smaller literature of the effects on crime (compare, for instance, Galiani et al. 2011) and some evidence on how conscription affects the market for education or the labor market, for instance provided by Card and Cardoso (2012) and Maurin and Xenogiani (2007). Considerably less, however, has been said about the effects of CMS and its alternative services on the market sectors conscripts were working in. To the best of my knowledge this is the first paper providing evidence on this aspect of drafting.

1.3 Institutional Background

1.3.1 Compulsory Military Service in Germany

Over recent decades, many European countries have changed the recruitment of military personnel. Whereas most countries pursued a policy of CMS for at least some time

[3]Hospitals' labor markets or input and patients well-being have also been considered in the broader context of labor market regulation (e.g Propper and Reenen 2010), hospital competition (e.g. Propper et al. 2004; Cookson et al. 2013) and market consolidation (Ho and Hamilton 2000). Yet, these and similar studies in general do not consider explicitly the structure of the hospitals labor force and thus can be seen as a different area of research.

[4]Interestingly, empirical evidence on educational outcomes suggests that conscription may lead to higher numbers of university graduates due to incentives to avoid the draft (Bauer et al. 2014).

during the last century, currently only few European countries (e.g. Austria, Finland and Greece) rely on drafting. Most of the states abolished or suspended military drafting several years ago (e.g. Netherlands (1997), France (2001), Italy (2005), Croatia (2008) and Poland (2010)) and now run professional armies[5]. Germany followed this trend when it indefinitely suspended the CMS in 2011 after more than half a century of mandatory military service.

Until 2011, according to the CMS Act introduced in 1956 (*Wehrpflichtgesetz*), every German male aged 18 to 23 was drafted for a period between 6 to 18 months[6]. In 1960 the alternative possibility to perform community service was introduced. By the Civilian Alternative Service Act (*Zivildienstgesetz*) young males were given the possibility to serve in a position "*for the common good*" (community service[7]) as part of a conscientious objection of military service. Young men performing community service had to serve between 6 and 20 months. There were a variety of options where to perform community service, yet, as can be seen in Figure 1.1 most people worked in care or care related jobs. Whereas in the 60s and 70s only few draftees chose the non-military service, it became rather common to object military service in the 1990s and 2000s. In 1961 only 574 young men were drafted for community service, but numbers went up to 135,924 decades later and remained on high levels ever since, stabilizing at around 100,000 about 10 years prior to the suspension (compare Figure 1.2). As a consequence, the share of draftees performing community service increased substantially over the decades and reached numbers above 50 percent 10 years before the suspension as can be drawn from Figure 1.3. Two developments are likely to have contributed to this increase in take up rates. Changing societal attitudes towards alternative services and the fact that the length of time young men had to serve under either of the options converged progressively. Whereas for most of the time (1973 to 2004) young men had to serve for a longer period if they opted for non-military service jobs, both services lasted 6 months in the last years before the cessation of drafting[8]. This includes the entire observational period of this paper. Thus, eventually there was no incentive to opt for military service besides the

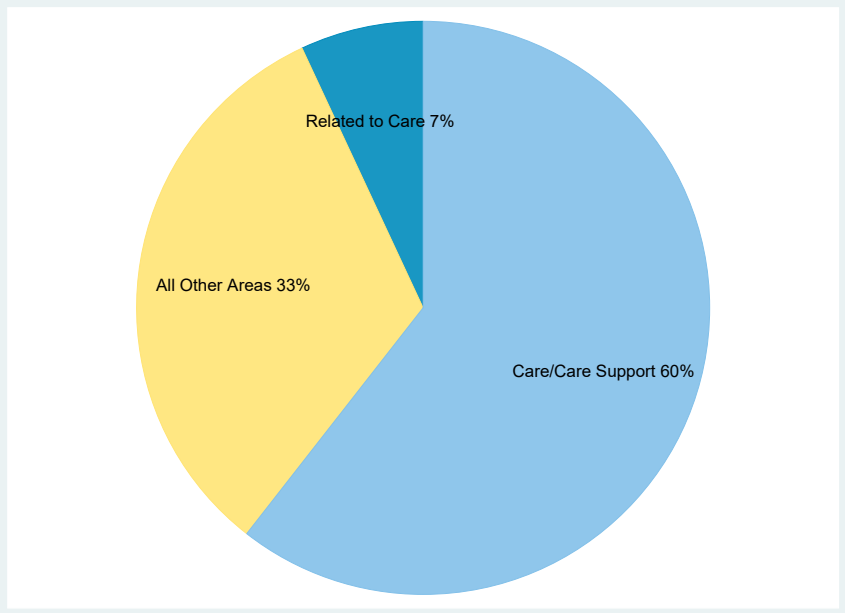
[5]Compare CIA Worldfactbooks: Central Intelligence Agency. "World Factbooks" https://www.cia.gov/library/publications/the-world-factbook.html, Central Intelligence Agency, n.d., Web, last accessed on 20 Oct. 2017

[6]The number of months depended on the year the recruit turned 18 in. By tendency, recruits were conscripted for more months in earlier decades of the draft.

[7]This paper uses the terms community service, non-military service, and alternative service interchangeably. All of these terms refer to the type of service falling under the Civilian Alternative Service Act (*Zivildienstgesetz*).

[8]Despite the fact that according to §12a(2) of the Civilian Alternative Service Act the alternative service was not supposed to take longer than the military service, for most of the time, the alternative service took between one and five months more than the military service. This was officially justified by the fact that draftees in the armed forces had to work longer hours, for instance, when participating in field maneuvers.

Figure 1.1: Areas of Employment of People Performing Community Service in 2010



Source: Federal Agency for Family and Civil-Societal Tasks, Own Calculations

content of the tasks performed. In the years prior to the suspension, the majority of the draftees preferred to perform non-military service. While, in 2010, less than 70,000 young men agreed to serve in the armed forces almost 80,000 were drafted for substitute services. This system was in place and enacted for 50 years until in December 2010 the German federal cabinet voted for the indefinite suspension of the CMS starting in July 2011.

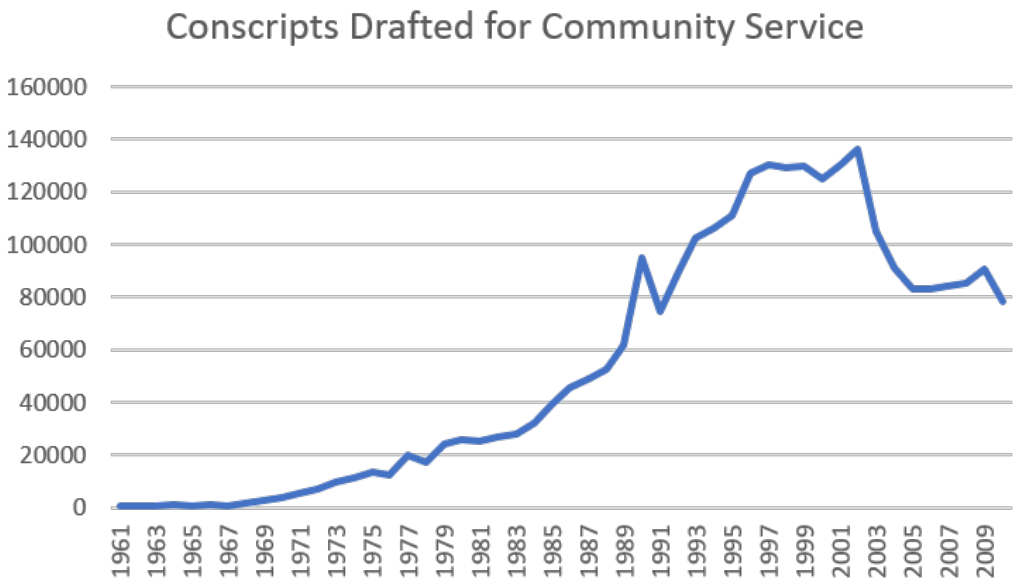
1.3.2 The German Hospital Sector

The German health care sector is highly regulated by legislation on federal and state level. This applies specifically to the hospital sector, which is subject to supply planning and a dual system of financing mainly provided by federal states and health care providers.[9]. According to these regulations public institutions are responsible for covering the capital costs of hospitals that are subject to the hospital requirement plan (Krankenhausbedarfsplan). The hospital requirement plan has to be provided by the federal states. It includes the number of hospitals and beds required to cover the medical needs of the states' population, including investments, building occupancy expenses, and acquisition costs of large medical equipment[10]. For this the federal states decide on the capacity requirements (e.g.

[9]Details of this funding scheme are laid down in §17b of the Hospital Financing Act (Gesetz zur wirtschaftlichen Sicherung der Krankenhäuser und zur Regelung der Krankenhauspflegesätze (KHG)), in the Hospital Remuneration Act (Krankenhausentgeltgesetz (KHEntgG)) and in the case rate agreement of the self-governing partners (Fallpauschalenvereinbarung der Selbstverwaltungspartner).

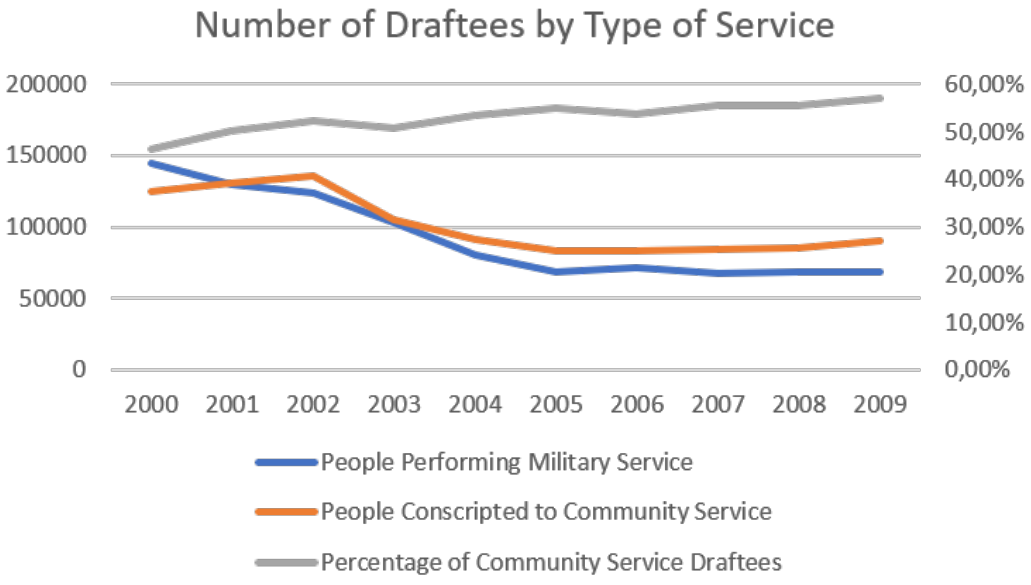
[10]Large medical equipment refers to all equipment that, according to law, can be written off over the course of at least 3 years.

Figure 1.2: Community Service Time Series



Source: Federal Agency for Family and Civil-Societal Tasks, Own Calculations

Figure 1.3: Time Series of Draftees



Source: Federal Armed Forces, Federal Agency for Family and Civil-Societal Tasks, Own Calculations

numbers of hospitals, location, specialization, emergency services or beds) a year ahead in the hospital requirement plan and provide monetary resources accordingly.[11] Overall, there is increasing cost pressure as financial resources provided by the states decreased by 30 percent since the 1990s (compare Destatis 2016).

A similar pattern of prospective planning applies to the running costs of the hospital, which are covered by health insurances based on a Diagnosis Related Group (DRG) reimbursement system. DRG is a classification scheme that serves as a basis for a lump sum billing system. Based on main and secondary diagnosis, procedure codes, and demographics each case is classified into an ICD-10 code[12] for which there are fixed reimbursements to the hospital[13]. The budget, type, and number of services are negotiated yearly between hospitals and health insurance providers[14] and there is a refund system between hospitals and health insurances if hospitals deviate from the plan. This means that hospital sponsors have to refund part of their earnings from the case based payments to the insurance companies, if they provided more treatments during a year than initially agreed on and vice versa. The rationale behind that is that hospitals are hardly able to adapt their (prospectively determined) labor and capital input within the current year and therefore should not have (much) more costs, regardless of the number of cases taken care of. Despite the fact that hospitals included in the supply plan have to take and treat patients "within the scope of their performance requirements and abilities"[15] [translation by author] this provides a strong incentive to stick to the initially agreed terms. The negotiation typically takes place in the beginning of the prospective year and can last until the end of the year. During our period of observation, however, negotiations were finished before the suspension of military drafting was discussed and decided on, leaving hospitals no scope to incorporate the policy change in their prospective planning before

[11] Additionally, there is the possibility for the hospital sponsor to apply to the regarding state for extra non-recurrent investments, for instance, if refurbishments or re-buildings are necessary.

[12] International Classification of Diseases, 10th revision

[13] There has been a worldwide trend towards prospective reimbursement systems. DRGs are at the center of this development. Originally introduced to standardize patients and thereby hospitals' services they are vastly used as an informational basis to rate and reimburse hospitals' services now, for instance, since 1984 by Medicare in the USA or since 2004 in all German hospitals, except for mental hospitals (Breyer et al. 2013). It is the aim of DRGs to classify patients into homogeneous cost groups. Therefore, patients are first assigned to a major diagnostic category, which is usually defined by the affected body part, and then further classified into DRGs based on their case and treatment. Apart from the diagnosis, complications and the undertaken procedures also are considered when determining the DRG. Therefore, precisely the system in place is not a purely diagnose-based classification and reimbursement system, but also includes fee for service and cost reimbursement elements (Breyer et al. 2013).

[14] Precisely, health insurance providers that insure at least 5 percent of the patients in the regarding hospital are involved in the negotiations.

[15] This obligation can be deduced from §108, Abs. 1, 109 SGB V.

it was implemented[16]. Note, however, that this does not apply to the years after 2011, i.e. there was not much scope to adapt to the policy change in-between planing and implementing the change in the yearly budget negotiations but hospitals were absolutely able to negotiate and adapt their staffing in following years, i.e. our after period.

1.3.3 The Suspension of Military Drafting and Low-Skilled Labor in Hospitals

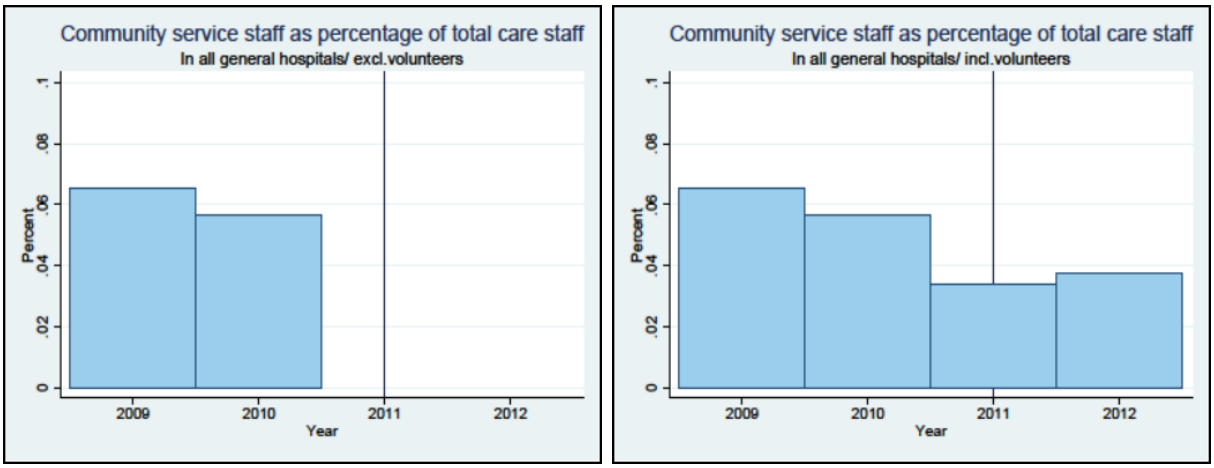
The suspension of military drafting was discussed in fall 2010 and already voted for in December of the same year. It was then implemented quickly, resulting in a large scale and fast reduction of low-skilled labor in hospitals. The last men were drafted in the following June such that by the end of 2011 there were no more men performing military or community service compulsorily. As a consequence, the number of people performing community service dropped from around 80,000 in the previous years to 0 by the end of 2011. For the average community service activity field this meant that one percent of their staff dropped out of their service fields within one year. Hospitals, medical prevention, and rehabilitation institutions were particularly affected by this policy change as until then around 67 percent of the people performing community service worked in hospitals and care homes (care/ care support) as well as ambulance services and patient transportation (areas related to care) (Figure 1.1). For these institutions people performing community service provided a considerable portion of the workforce. Due to the policy change the average general hospitals, under consideration in this paper, lost more than 6 percent of their care staff (Figure 1.4)[17].

Draftees in the health sector provided assistance to the care staff and patients directly. Typical tasks in hosptials were, for instance, changing and cleaning beds, assisting patients in getting around or simple care assistance like taking patients' temperature or feeding immobile people. The labor input of draftees by law was meant to be labor market neutral, implying that people performing social work were only allowed to be assigned to tasks that otherwise would not have been done by paid personnel. Yet, the fact that

[16]The suspension of drafting was voted for due to an initiative of the Free Democratic Party (FDP). This party had been elected into the government in federal elections 2009, when the government changed from a grand coalition between the Social Democratic Party (SPD) and the Christian Democratic Union (CDU) to a coalition between CDU and FDP. If hospitals anticipated this controversially debated policy change to be implemented solely by the fact that the liberal party had been voted for and adapted to it before the policy change was actually enacted, it would attenuate any effect on personnel or patients. One should, however, be able to observe changes in the input structure of hospitals from 2009 to 2010 then. Neither descriptive evidence, provided in the data section below (1.4.2), nor results provided in 1.5 show any evidence corroborating this view.

[17]Figure 1.5 depicts people performing community service in hospitals with different types of ownership.

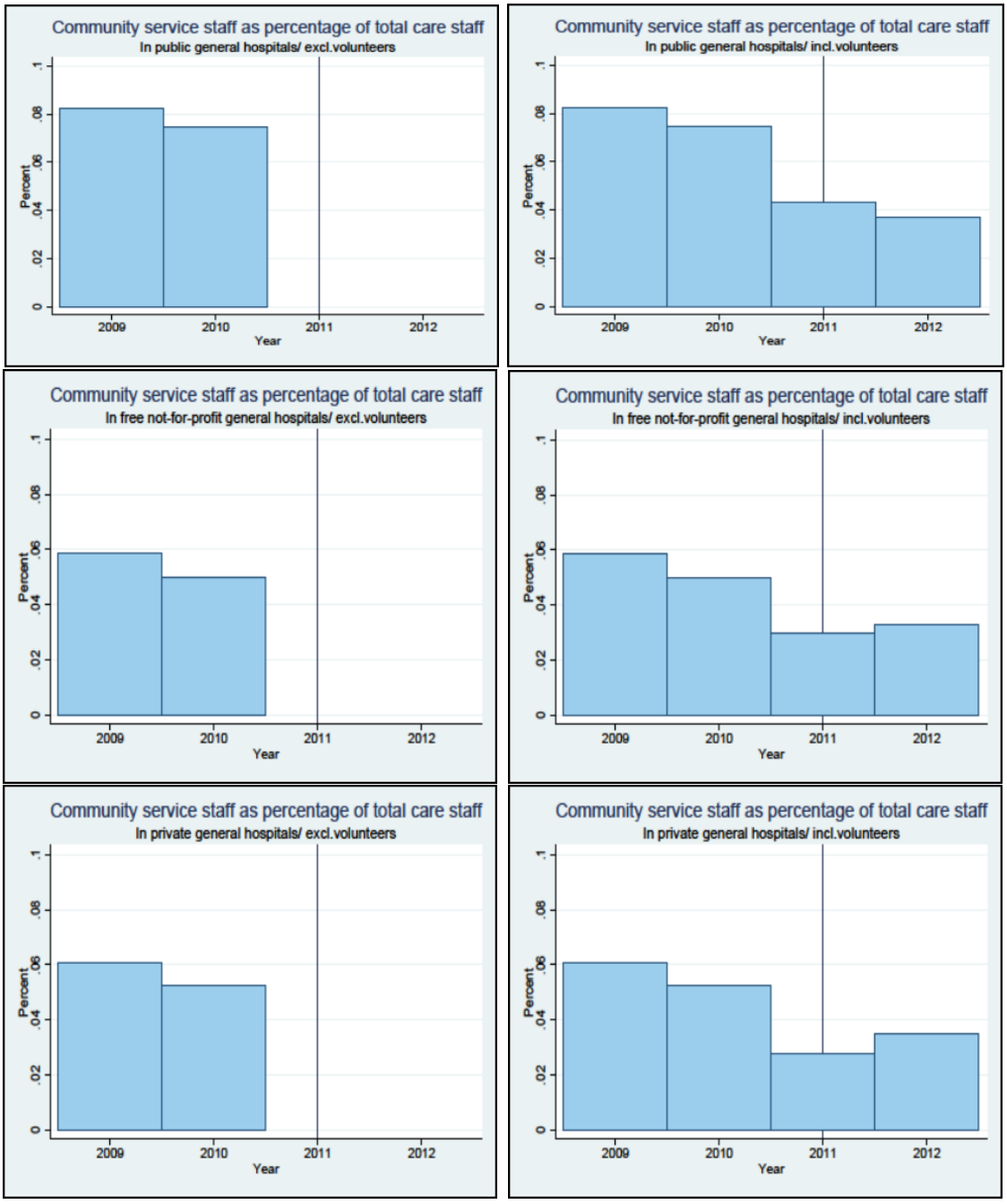
Figure 1.4: Drop in People Performing Community Service in General Hospitals



hospitals did not have the duty to hire people performing community service and at the same time had to pay for draftees' material if they did so makes it implausible that people performed community service in hospitals without any valuable contribution. This rationale is clearly supported by studies on behalf of the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (Beher et al. 2002: chapter 21.1). Even more so, the provider of jobs for draftees expressed their concerns regarding the quick suspension.
Thus, to account for the sudden loss of labor input for health care providers, previously existing voluntary services (Youth Volunteer Service, *Jugendfreiwilligendienst*) were extended by the government in the wake of the cessation of drafting. On top of that, more opportunities to perform voluntary services were introduced (Federal Volunteer Service, *Bundesfreiwilligendienst*). Comparable to community service, these services were accessible to volunteers in all states, yet, for a broader age group and with an even broader set of institutional partners. Still, people participating in these programs only accounted for less than half of the previous employees in hospitals (compare right panel of Figure 1.4), i.e. even when taking this into account, as the empirical analysis of this paper does, the policy change still caused a considerable drop in labor input. The drop is also rather similar across the different types of hospital ownership (compare right panel of Figure 1.5). Neither the people under the drafting system nor participants of the Youth Volunteer Service Program were offered any comprehensive job specific training before their employment[18]. Similar regulations applied to the community service. Therefore, draftees and volunteers can be regarded low-skilled with respect to the tasks performed in a hospital.

[18]Volunteers participating in the Federal or Youth Volunteer Service are obligated to participate in 25 training days per year of service, but these seminars only prepare participants on a general level and facilitate the exchange of experiences with other volunteers.

Figure 1.5: Community Service in General Hospitals by Type of Ownership



1.4 Empirical Analysis

1.4.1 Theoretical Considerations

To understand how hospitals are affected by the suspension of drafting theoretically, it is helpful to consider a hospital production function. The previous literature has shown that hospital production can authentically be described by versions of a Cobb-Douglas production function[19], compare, for instance, Jensen and Morrisey (1986) for one of the earliest references. The suspension of drafting led to a drop of people performing community service, i.e. the amount of labor input decreased. Standard theory implies that the effect of the policy change on hospitals would be a decrease in hospital production, i.e. the empirical analysis would show the significant drop in people performing community service induced by the policy change, no further effect on labor input, no effect on capital input and a decrease in production[20].

However, it is natural to assume that hospitals also respond to the policy change. Let us consider two different types of responses according to the Cobb-Douglas function. These cases should be viewed as theoretical cases for illustration. It is, of course, possible that adjustments are made in several dimensions at the same time.

Case 1: Hospitals adjust their labor input.

Hereby, three sub-cases can be distinguished. (a) The drop in people performing community service leads to a drop in overall labor input. One possible scenario would be that hospitals were not able to react to the exogenous shock in terms of staffing enough to compensate exactly for the drop. All other things equal, the initial reduction in labor input then leads to a reduction in hospital output. Empirically, we would expect to observe no or small positive changes in staffing (besides the negative effect on people performing community service, of course) and a negative effect on hospital output measures.

(b) It could be possible for hospitals to hire exactly the amount of human capital they lost

[19] For example:

$$f : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+,\ f(x_1, x_2) = Ax_1^a x_2^b, \quad 0 < a, b < 1,\ a + b = 1 \tag{1.1}$$

where $f(x_1, x_2)$ denotes hospital output, x_1 denotes capital input and x_2 stands for labor input. The parameters a and b show how the output reacts to changes in input and $A \in \mathbb{R}_+$ can be seen as the scale of productivity. For the qualitative considerations here it is not necessary to analyze these factors in more detail. The production function yields the following partial derivatives

$$\frac{\partial}{\partial x_1} f(x_1, x_2) = aAx_1^{a-1}x_2^b, \quad \frac{\partial}{\partial x_2} f(x_1, x_2) = bAx_1^a x_2^{b-1},$$

whereas $A, a, b, x_1, x_2 > 0$ imply that both derivatives are strictly positive.

[20] This, of course, excludes the possibility that people performing community service did not contribute any valuable labor input to institutions production. Then output would not be affected at all. Reasons why this theoretical case can be excluded were already outlined above.

through the policy change. We would then observe an increase in at least one staff group and no effect on output. In most cases conscripts did not have any or only few weeks of training to prepare for their community service jobs. As such, they can be considered low-skilled. If hospitals were to adjust labor input it is plausible that they did so by employing other low- or medium skilled workers. The employees closest in skills would be volunteers participating in the federal and youth volunteer program, who also did not have any formal training regarding their activity fields. Unlike in other countries, where nursing requires a university degree, nurses in Germany undergo a 3-year vocational training. Despite the fact that this training is considered challenging and includes a rather high amount of theoretical instruction, nurses are primarily concerned with supporting the therapies prescribed by doctors. As such, they can be considered medium skilled and could also be suitable substitutes for conscripts' labor. A similar reasoning applies to nursing apprentices. Economic theory would predict an increase in labor input by these personnel groups if hospitals were to substitute.

Doctors, on the other hand, rather function as complements to low(er)-skilled staff in hospital production. A medical treatment in a hospital usually involves physicians doing check-ups and prescribing or carrying out treatments and care staff supporting their therapies. If doctors are complements to other skill groups in hospitals, we should observe a change in the same direction as with the overall input from low and medium skilled personnel. In the same sense medical equipment (part of the capital input of hospitals) is complementary to doctoral staff, because physicians have to approve or prescribe their usage or operate it.

(c) It could also be possible that hospitals hire more human capital in response to the policy change. Due to a shortage of volunteers, for instance, hospitals could replace each draftee by a nurse and consequently would operate with the same head count but more able teams, predicting positive effects on production.

Case 2: Hospitals adjust their capital input.

As can be derived from the production function above, capital and labor can be used in different ratios to produce the same hospital output. It is therefore possible that hospitals respond to the initial drop in labor input by increasing their capital input. Applying a similar reasoning as above three main cases can be distinguished here as well. Hospitals could (a) not respond at all, (b) respond a little bit but not fully offset the effect of the labor decrease or (c) over-compensate the labor input, possibly due to capital investments being discrete (for instance, a software that assists in monitoring patients replacing the work of more draftees than lost). All other things equal, each increment in capital has a positive effect on hospital production. The overall effect on production in these three cases, however, depends on the interplay between capital and labor and is eventually an

empirical question.

1.4.2 Data Source: The German Hospital and Patient Statistics

Shedding light on the theoretical mechanisms just outlined requires very comprehensive data on hospital's in- as well as outputs. These data are contained in the German Hospital and Patient Statistics (*Krankenhausstatistik I-III*), an administrative data set collected by the statistical offices of the German federal states, covering all hospitals of the country[21] and the vast majority of their patients every year[22]. This statistic serves as the planning basis for the public authorities involved in financing of hospitals and clinics. As such, every hospital and clinic is obliged by law (§28 KHG) to provide detailed information on their organizational structure, staff, equipment and costs. Overall, the Hospital and Patient Statistics 2009 to 2012 include between 2017 and 2084 economic entities[23] and a total of 75 million patient cases[24].

For the analysis in this paper I restrict the sample to general hospitals, because prevention and rehabilitation clinics as well as pure day or night clinics are structurally very different from general hospitals. They have rather fixed durations of stay and very low mortality rates, for instance. Also, they were not relying on people performing community service as much. Furthermore, I exclude from the analysis general hospitals with an average duration of stay below one day, hospitals with no directly employed doctors or nurses and hospitals where the average cost for nurses are higher than the costs for doctors. This excludes highly specialized clinics such as beauty clinics or hospitals that disproportionally rent out beds to resident physicians, for instance, for smaller surgeries. To analyze the effect of the suspension of CMS I will make use of three main sets of outcome variables (hospital outcomes and labor and capital input).

Measuring **hospital's outcomes** is inherently difficult because hospitals produce a mul-

[21] Hospitals by German law (§2(1) KHG (Hospital Financing Act)) are "institutions that diagnose, cure or mitigate diseases, conditions or physical injuries or assist at birth by doctoral or nursing support that accommodate and keep the persons to care for" [translation by author] i.e. by definition the observational basis is rather broad, including prevention and rehabilitation clinics, which are excluded from the analysis.

[22] Only very small entities with less than 100 beds are exempted from the obligation to provide information on their patients.

[23] For the sake of simplicity in this paper one economic entity is referred to as a hospital. Note, however, that one economic entity can include several hospitals or different sites (Information by Data Specialist Department of Statistisches Landesamt Hessen, as received via email by Dr. Peter Gottfried on Aug, 2, 2017). The number of hospitals decreases slightly over time because of increasing market concentration due to merging, i.e. initially individual hospitals merge to one larger hospital with several sites. These are then surveyed under one hospital identifier. There is also a slight reduction by hospital closure.

[24] Note that, although there is information on transferals and shifts of patients, it is not possible to link the data of previous stays to current stays of re-admitted patients.

tidimensional and to a certain degree subjective outcome: health. They also stand by to fulfill an optional demand, i.e. they withhold a certain capacity, for instance, for emergency treatment (compare conclusion 9.1, pp.355 in Breyer et al. 2013). The value of this good can also be rather subjective and particularly hard to quantify especially when not utilized. To measure hospital outcomes the empirical analysis will make use of the overall number of cases treated, numbers of in- and out-patients, numbers of death as well as rates, and likelihoods to be discharged to another hospital instead of being able to go home after treatment. While there is a much bigger variety of potential output measures (e.g. failure to rescue rates or case progression severity), these are seen to be the most objective outcome measures[25].

An analysis of the development and interaction of overall case numbers as well as numbers of in- and out-patients treated allows us to identify one important potential way of reacting to the policy change if original staffing levels cannot be retained. The number of out-patients refers to registered cases for which hospitals reported a duration of stay of a day or less in general hospitals[26]. In-patients are identified by a duration of stay of two days and above. Duration of stay reports the average duration of stay of all in-patients in the final sample in a regarding hospital. Due to the diagnoses related case based refund scheme hospitals have a strong incentive to discharge patients as soon as possible, i.e. as soon as a patient reaches the necessary level of health to be discharged. A longer duration of stay means that hospitals need more time to reach that level indicating adverse effects on patients' health. Mortality is measured in absolute numbers of in-hospital death incidences as well as in share of people dying in hospital (=case fatality rate). Mortality is the most severe clinical outcome, but is also commonly used as a quality indicator for hospitals. One clear advantage of this outcome is that it is robust to different coding behavior of hospitals (Wissenschaftliches Institut der AOK 2007: p.29) and thereby easily comparable across hospitals[27].

The main dependent variables to measure **labor input** are yearly average full-time equivalences (Full-Time Equivalences (FTE)) of all low, medium and high-skilled staff. Yearly average FTEs are calculated based on the working hours of all contracts over the course of the year. A full-time employed nurse working from January until March would, for instance, be reported as 0.25 FTE. The same applies to all other trained groups consid-

[25] A detailed discussion of potential outcome measures for analyzing hospital production can, for instance, be found in Cook et al. (2012).

[26] This is done because registering a patient with a duration of stay of one day is very common even if the actual hospital stay takes less than 24 hours and does not include staying over night.

[27] This can be of relevance because hospitals record their own data and know that this data can be used for qualitative evaluation. Therefore, despite the fact that there are strong incentives to report data correctly as to being paid accordingly, there might be an incentive to down play adverse events like complications which would bias indicators like failure to rescue rates.

ered[28]

To provide a complete picture of care staff's labor input, 3 more measures of labor input are defined. *Nurses per bed* refers to yearly average FTEs of care personnel per installed hospital bed and reflects the potential capacity care personnel need to cover. *Workload of nurses in terms of case numbers* relates the number of cases per hospital and year to the number of FTEs. This is of interest because every case causes a certain amount of time consuming registration and discharge paper work and therefore causes a workload different from an in-hospital care day. Finally, *workload of nurses in terms of care hours* is an indicator of how many care hours (reported care days x 24) were performed per actual working hours of a yearly average full-time equivalent (FTE x 220 [working days per year] x 8 [hours]), which provides us with a measure of how strenuous an actual workday is. The same measures are defined for all other skill-groups.

The main dependent variables for analyzing **hospital's capital input** are hospital's material costs and overall expenditure in Euros per year. Material costs is a summary measure for all capital input of the hospital[29]. They include various types of costs related to the treatment of the patient. In particular, medical consumables like drugs, bandages, instruments, therapeutic appliances, blood and plasma are contained in this measure. Any form of expenditure on labor is excluded. Additionally, this paper provides evidence on overall hospital expenditure. Here, a summary measure for all costs of the hospital that are not covered by public authorities, including material costs, staff costs, training costs just as well as costs for hospital administration is constructed. As labor expenditures are included these costs provide a complete picture of hospital costs. Additionally, this paper provide more detailed evidence on how expenditure on staffing changed in response to the policy change. Costs for the single staff groups are split up in eleven subcategories in the original data (including, for instance, doctoral services, care services, medical-technical services, and administration). This paper considers doctoral services, care services, and expenditure on staff training in more detail. All of these are measured in yearly sum per hospital grouped by staff category. Considering these measures provides interesting additional insights into staffing beyond contracted labor input.

As additional **control variables** hospital characteristics like type of sponsor (public, private, non-statutory welfare services), type of hospital (i.e. university hospital, hospital forming part of the German hospital plan, hospital with provision contract or others), and share of apprentices in total staff are used. Hospital size is controlled for by number of

[28] Apprentices in contrast are converted with a factor of 9.5 to 1 or 6 to 1, depending on their type. People performing community service are converted 1 to 1 as they were employed full-time.

[29] Capital costs which are covered by state authorities (e.g. large medical equipment, buildings) are excluded because they are subject to prospective planning and should react differently, if at all, to changes in the labor structure.

installed beds. Following the standard approach in the literature (compare, for instance, Breyer et al. 2013) these are summarized in 5 categories (less than 75, 75-124, 125-249, 250-499, 500 or more). Additionally, regional characteristics are accounted for by controlling for the degree of regional agglomeration by considering 8 categories ranging from densely populated urban areas to rural county, lightly populated. This is relevant because descriptive evidence has shown marked differences between treatment prevalence across counties. Furthermore, share of female patients, share of patients aged 75 and above, and share of patients undergoing surgery are included to control for hospitals patient mix. Further details on the data set, sample, and the definition of each variable are contained in the data appendix.

Data on yearly averages for additional years was drawn from the basic data on hospitals published by the Federal Statistical Office (Destatis 2017). Complementary data on draftees as well as specific general information on military service were obtained from the Federal Armed Forces (*Bundeswehr*). Further supplementary data on community service, as, for instance, numbers of people performing community service or types of tasks performed, were obtained from the Federal Agency for Family and Civil-Societal Tasks.

Table 1.1 shows descriptive statistics for the sample of general hospitals separately for the pre-reform (2009/2010) and the post-reform period (2011/2012). The descriptive statistics clearly show the drop in people performing community service in hospitals. Whereas there were 11.5 people performing community service per hospital before the policy change, there were none in 2011 and 2012. The descriptives also show that the staffing situation in hospitals improved over time. In the after reform period, there have been more nurses, more specialized care staff, doctors, and apprentices per hospital. At the same time, the number of hospitals decreased. The higher staffing levels could, therefore, also be related to market concentration, i.e. there are fewer, but bigger hospitals and need not indicate that people performing community service have been replaced. The fact that all measures of nurses workload remain rather constant corroborates this view.

1.4.3 Empirical Approach

I estimate the effect of an exogenous change in the amount of low-skilled personnel on a range of hospital and patient outcomes. The empirical strategy of this paper exploits the fact that not all hospitals employed people performing community service. Hospitals that employed people performing community service experienced an exogenous change in their staff structure due to the policy change and will be included in the treatment group. All other hospitals are assigned to the control group. The main specification then compares hospitals that did experience an exogenous drop in people performing

Table 1.1: Summary Statistics

	(2009-2010)		(2011-2012)	
	Mean	Std. Dev.	Mean	Std.Dev.
Input				
Community Service (Staff No.)	11.5	16.3	6.9	11.4
Nurses (FTE)	178.2	226.9	187	236.7
Nurses per Bed	0.57	0.21	0.59	0.22
Workload Nurses (Care Hours)	6.9	3	6.7	3.1
Workload Nurses (Cases)	68.3	34.5	69.7	35.9
Other Medical Staff (FTE)	141.9	271.2	152.7	286.4
Doctors (FTE)	81.5	141.2	88.8	149.4
Apprentices (FTE)	11.3	31.3	12.5	29.7
Output				
Number of Out-Patients	1,696	2,083	1,906	2,276
Number of In-Patients	9,536	10,406	10,043	10,850
Share of In-Patients	0.86	0.08	0.85	0.09
Number of Beds	293	306	300	315
Number of Care Days	81,803	91,569	83,656	93,844
Average Staying Duration	8.7	11.4	8.7	12
Share of Patients Dying	0.0224	0.0162	0.0211	0.0153
Costs				
Total Expenditure	46.7	1.4	52.1	1.5
Material Costs	18.3	0.6	20.3	0.6
Expenditure Care Staff	8.8	0.2	9.6	0.2
Expenditure Apprentices & Training	0.3	0.01	0.4	0.01
Expenditure Doctoral Staff	8.3	0.2	9.7	0.3
Controls				
Share of Patients Undergoing Surgery	0.3021	0.231	0.3012	0.2331
Share of Female Patients	0.542	0.0919	0.5383	0.0898
Share of Patients Aged 75 and Above	0.2765	0.125	0.2886	0.1269
Observations Public	795		808	
Observations Not-for-Profit	1,062		1,010	
Observations Private	1,251		1,203	
Observations Total	3,108		3,021	

Notes: The table reports summary statistics of the sample of general hospitals and their patients in Germany over the period from 2009 to 2012. The first two columns report mean and standard error for main outcome and control variables for the pre-policy period, the last two columns report the same statistics for the post-policy period. Staffing is reported in terms of numbers (regardless of type or scope of employment relationship) and yearly average full-time equivalences. Nurses per bed refers to yearly average full-time equivalences per hospital bed and workload of nurses (care hours) is an indicator of how many care hours (reported care days x 24) were performed per actual working hour of a yearly average full-time equivalent (FTE x 220 (working days per year) x 8 hours). The workload in terms of cases relates the number of patient cases per hospital and year to the number of full-time equivalences of nurses. Number of out-patients refers to numbers of registered cases for which hospitals reported a staying duration of a day or less in general hospitals. Note that registering a patient with a staying duration of one day is very common even if the actual hospital stay takes less than 24 hours and does not include staying over night. Number of in-patients refers to cases with a staying duration of above one day. Average staying duration is calculated on the basis of all patients in the final sample. The same applies to share of people dying, female patients and patients aged 75 and above. The cost measures refer to average costs per year and hospital.

Source: German Hospital and Patients Statistics (2009-2012)

community service to those that did not in a differences-in-differences approach[30]. In 2010, the year prior to the policy change, 482 hospitals did not report any people performing community service[31]. These hospitals did not experience any changes due to the policy change directly, as their staff structure was not affected by the suspension of military drafting. Consequently, these hospitals serve as the control group. The model can be written in the following way:

$$y_{ht} = \beta_0 + \beta_1 above_h + \beta_2 after_t + \beta_3 (above_h \cdot after_t) + \lambda' X_{ht} + \varepsilon_{ht} \tag{1.2}$$

where $above_h$ is equal to one in all years t if the number of people performing community service in a given hospital h is greater than zero in 2010 and zero otherwise, splitting the sample into treatment and control group. The coefficient then indicates the average difference between the two groups. $after_t$ is a dummy variable equal to one after the suspension of military drafting. As the German Hospital Statistic samples all hospital information as matters stand on December, 31, 2011 and 2012 can both be regarded as observations after the policy change. This however, does not apply for yearly average full-time equivalences and yearly costs, which is accounted for by solely considering 2012 as the after period in the regarding specifications. The coefficient of $after_t$ reflects the average difference between the before and after period, i.e. the general time trend. The interaction term between $above_h$ and $after_t$ then introduces the differences-in-differences structure to the model and its coefficient reflects the change in differences between the two groups caused by the policy change. β_3 therefore is our main parameter of interest.
I also include a set of control variables, contained in X_{ht}, by controlling for observable differences these variables reduce the residual error and improve the precision of the point estimates. Following the previous literature on hospital production (for instance Herr 2008; Herr et al. 2011; Steinmann and Zweifel 2003), the main specification controls for hospital characteristics, that have been shown to significantly impact hospital efficiency, such as hospital size (in bed number categories), legal form, sponsor, and share of apprentices. Furthermore, to control for hospitals' patient case mix, the share of females among the patients, the share of people aged 75 and above as well as the share of patients that underwent surgery are included in this model. Both, including hospital fixed effects as

[30] In principle, this setting would also be feasible for the application of a regression discontinuity or regression kink design. Unfortunately, the nature of the data at hand (only a few years and yearly averages for most outcomes of interest, covering up changes around the discontinuity) does not permit the implementation of such an approach.

[31] This includes hospitals that reported 0 people performing community services as well as missing values. Hospitals are not only obligated by law to report staff, they also have clear incentives to report all people performing community service, because these peoples' salaries were partly paid by public authorities. Therefore, I also include hospitals that reported a missing for people performing community service in the control group, assuming these hospitals simply did not report zero values.

well as adding the set of control variables just discussed - eventually accounting for the differences across hospitals - are common in the literature. The advantage of including the control variables is that the influence of the single factors becomes transparent and thereby sheds some light on the underlying mechanisms.

Descriptive evidence on the regional distribution of disease prevalence as well as treatments indicates substantial regional variation in the quantity of health care services performed[32], compare, for instance, Nolting et al. (2011). To account for this correlation within regions, standard errors ε_{ht} are clustered at the nuts-2 level, leaving us with 39 regional clusters[33].

Key to the identification strategy of this paper is the parallel trend assumption between hospitals that employed people performing community service and hospitals that did not. For the identification strategy to be credible, hospitals of the treatment and control group need to share the same time trend in potential outcomes (compare, for instance, Angrist and Pischke 2009). To corroborate this assumption a graphical test for all sets of dependent variables (hospital outcomes, staffing, and costs) can be found in Figure 1.6. The graphs depict all of the outcome variables (y-axis) for the four years (x-axis) observed in the data set, including the two pre-policy years 2009 and 2010. As can be seen in the graphs only minor differences in changes from 2009 to 2010 can be observed. Further descriptive evidence on the pre-treatment period can be drawn from Table 1.2, which reports the pre-treatment outcomes by years. Ideally, of course, one would observe a much longer pre-treatment period.

[32]Interestingly, the reasons for this regional correlations are largely unclear. Explanatory approaches include, but are not limited to, regional differences in indication, possibly due to regional differences in training and experience levels of doctors, as well as differences in availability of and access to specialists. Previous attempts to correlate these explanations to procedures however did not manage to establish clear correlations. For a detailed overview of several regional descriptive indicators based on the German Hospital Statistics and several other German data sources refer to Nolting et al. (2011).

[33]This improves upon typical numbers of clusters, yet, might still be problematic (Bertrand et al. 2004; Abadie et al. 2017)

Figure 1.6: Key Variables Over Time by Treatment and Control Group

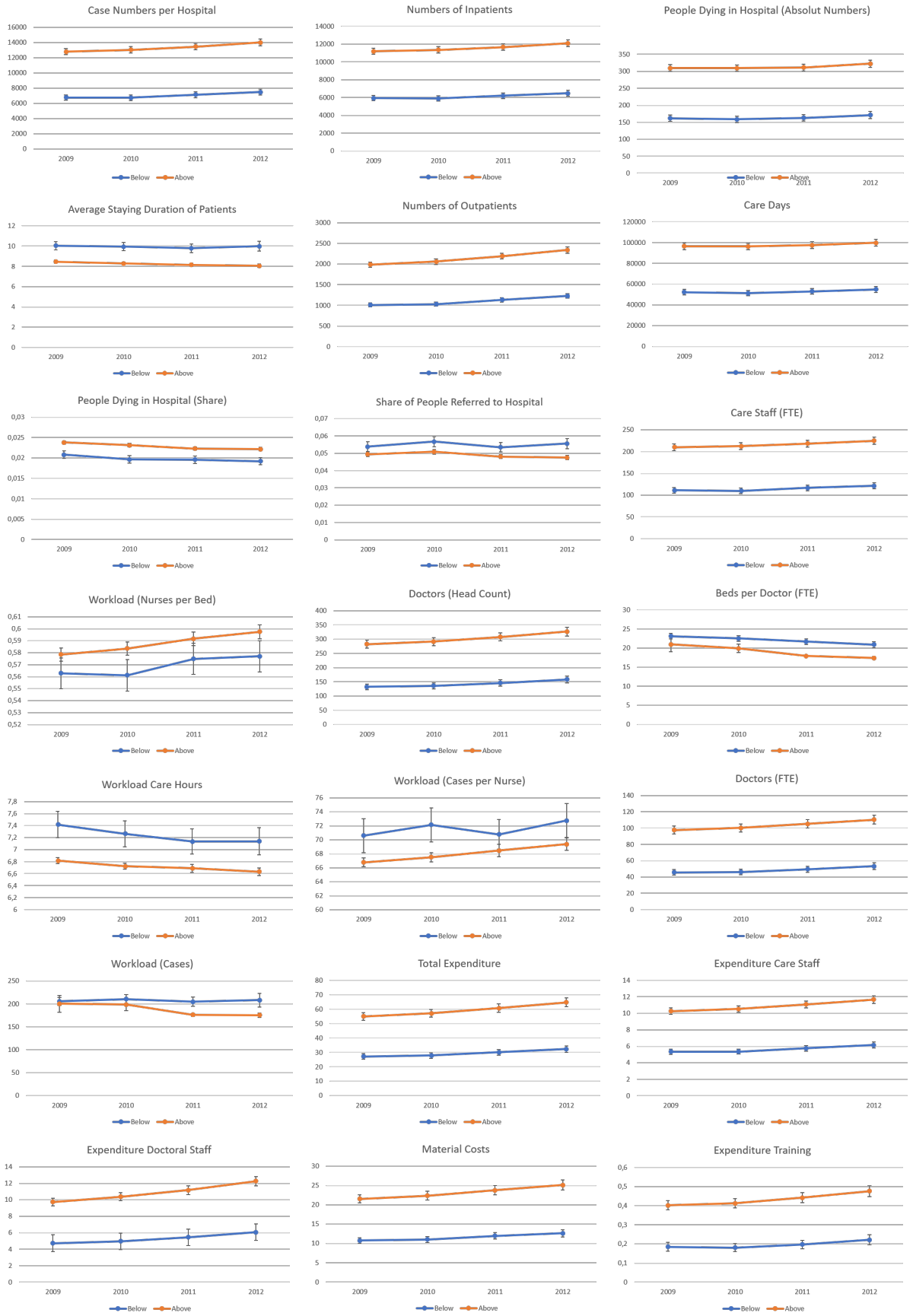


Table 1.2: Summary Statistics Pre-Treatment Period by Year

	2009				2010			
	Treatment		Control		Treatment		Control	
	Mean	Std.Dev.	Mean	Std.Dev.	Mean	S.E.	Mean	S.E.
Input								
Community Service (Staff No.)	13	0.6	7	0.8	11	0.5	0	0
Nurses (FTE)	210	7.7	110	6.4	212	7.8	110	6.1
Workload Nurses (Care Hours)	6.8	0.0	7.4	0.2	6.7	0.1	7.3	0.2
Workload Nurses (Cases)	66.8	0.6	70.6	2.4	67.5	0.7	72.1	2.4
Doctors (FTE)	98	4.9	46	3.4	100	5.0	46	3.5
Workload Doctors (Care Hours)	20.9	1.9	23.1	0.7	19.9	1.1	22.5	0.7
Workload Doctors (Cases)	200.2	18.6	205.3	8.9	198.5	13.3	210.3	9.7
Output								
Number of Out-Patients	1987	69.0	1009	62.3	2057	71.1	1029	63.8
Number of In-Patients	11201	347.0	5943	305.2	11367	353.4	5899	305.5
Number of Care Days	96406	3099.2	52252	2609.4	96299	3119.8	51267	2597.2
Average Staying Duration	8	0.2	10	0.4	8	0.1	10	0.4
Share of Patients Dying	0.024	0.0004	0.021	0.0010	0.023	0.0004	0.020	0.0009
Costs								
Total Expenditure	55.0	2.6	27.2	1.9	57.1	2.7	27.8	1.9
Material Costs	21.5	1.1	10.7	0.7	22.4	1.1	11.0	0.8
Expenditure Care Staff	10.3	0.4	5.3	0.3	10.5	0.4	5.4	0.3
Expenditure Apprentices & Training	0.4	0.0	0.2	0.0	0.4	0.0	0.2	0.0
Expenditure Doctoral Staff	9.7	0.5	4.7	0.4	10.4	0.5	4.9	0.4
Controls								
Share of Patients Undergoing Surgery	0.29	0.0057	0.30	0.0137	0.30	0.0057	0.32	0.0142
Share of Female Patients	0.54	0.0022	0.54	0.0059	0.54	0.0022	0.54	0.0056
Share of Patients Aged 75 and Above	0.28	0.0033	0.26	0.0072	0.29	0.0033	0.26	0.0072
Observations Total	1050		471		1061		482	

Notes: The table reports summary statistics of the sample of general hospitals and their patients in Germany over the period for the two pre-treatment years, seperately by treatment and control group. Staffing is reported in terms of numbers (regardless of type or scope of employment relationship) and yearly average full-time equivalences. Nurses per bed refers to yearly average full-time equivalences per hospital bed and workload of nurses (care hours) is an indicator of how many care hours (reported care days x 24) were performed per actual working hour of a yearly average full-time equivalent (FTE x 220 (working days per year) x 8 hours). The workload in terms of cases relates the number of patient cases per hospital and year to the number of full-time equivalences of nurses. Number of out-patients refers to numbers of registered cases for which hospitals reported a staying duration of a day or less in general hospitals. Note that registering a patient with a staying duration of one day is very common even if the actual hospital stay takes less than 24 hours and does not include staying over night. Number of in-patients refers to cases with a staying duration of above one day. Average staying duration is calculated on the basis of all patients in the final sample. The same applies to share of people dying, female patients and patients aged 75 and above. The cost measures refer to average costs per year and hospital.

Source: German Hospital and Patients Statistics (2009-2010)

1.5 Results

1.5.1 The Drop in People Performing Community Service

The identification strategy of this paper rests on the assumption that the suspension of drafting led to a significant drop in people performing community service in German general hospitals. As already shown descriptively and discussed above two things happened in 2011. On the one hand, drafting was suspended. On the other hand, to account for the sudden loss of labor input for health care providers, the previously existing voluntary service (Youth Volunteer Service, *Jugendfreiwilligendienst*) was extended by the government. On top of that, more opportunities to perform voluntary service were introduced (Federal Volunteer Service, *Bundesfreiwilligendienst*). Descriptively, people participating in these programs only accounted for half of the previous employees in hospitals (compare Figure 1.4[34]). Yet, for the empirical analysis it is crucial that, even taking these volunteers who are very similar to people performing community service into account, hospitals still experienced a significant exogenous shock in their low-skilled labor input. Table 1.3 provides evidence that this indeed was the case. Columns (1) and (3) each provide the pure differences-in-differences effect without including any controls and columns (2) and (4) provide results of the specification including the full set of control variables. As can be seen in columns (1) and (2) even including volunteers in the staff count for low-skilled labor a significant reduction in labor input can be observed. Columns (3) and (4) provide evidence on the development of low-skilled staff before the policy change, revealing a positive effect in column (3) and no effect on column (4). The positive effect on column (3), however, is only 0.075 percent of the actual effect induced by the policy change. It is important to note that all following results already take the increment in labor from volunteers into account, i.e. we only observe the effect of the "net" policy change.

1.5.2 The Drop in People Performing Community Service and Quantity and Quality of Hospital Production

The empirical analysis is proceeded by investigating how the drop in people performing community service affected quantity and quality of hospital production, in particular treated cases and objective treatment success. Table 1.4 reports results of differences-in-differences estimations of hospital outputs measured by four different indicators. The effects of the policy change on the number of patients treated each year are reported in column (1) and (2), the same numbers broken down by type of patient (in- vs. out-patients)

[34]Further evidence on all hospitals can be found in the appendix (Figure 1.7)

Table 1.3: Effects of Policy Change on Staffing: The Drop in People Performing Community Service

	2010 to 2012		2009 to 2010	
	(1)	(2)	(3)	(4)
Additional Change in Treated Hospitals	-2.215**	-2.662**	0.002***	0.001
(S.E.)	[0.951]	[1.155]	[0.000]	[0.000]
Hospital Controls	No	Yes	No	Yes
Regional Controls	No	Yes	No	Yes
Patient Mix Controls	No	Yes	No	Yes
Observations	2,580	2,580	3,083	2,580
R-squared	0.013	0.401	0.024	0.425

Notes: The table shows how the staffing with people performing community service responds to the policy change (columns (1)-(2)). As a robustness check it also shows changes in the pre-treatment period (columns (3)-(4)). Both specifications take the volunteers into account who stepped in for draftees. People performing community service and volunteers are counted in yearly average FTE. The first specification always shows the pure difference-in-difference estimator and the second specification includes the full set of control variables to gain precision. The full set of control variables includes dummy variables for public, private and free or not-for-profit providers, bedclass indicators, summarized in 5 categories (less than 75, 75-124, 125-249, 250-499, 500 or more), controls for the patient composition of hospitals (i.e. the share of females, the share of patients aged 75 and above and the share of people that undergo surgery during their hospital stay), dummy variables for the type of hospital (i.e. university hospital, hospital forming part of the German hospital plan, hospital with a hospital provision contract or other hospitals), the share and squared share of apprentices in total staff and regional characteristics (considering 8 categories ranging from "densely populated urban areas" to "rural county, lightly populated"). Shares are calculated based on staff and patients in the final data set. The sample is restricted to general hospitals with more than 100 beds. Standard errors are clustered at the nuts-2 level. *** p<0.01, ** p<0.05 and * p<0.1.

Source: German Hospital and Patient Statistics (2009-2012)

can be found in columns (3) to (6) and the overall number of days of care provided by the hospital can be drawn from columns (7) and (8). The key independent variable in all of the specifications reported in this table is the interaction term in equation (1.2). It equals one for treated hospitals after the policy change and its coefficient indicates the additional change a treated hospital experiences due to the policy change reported in the first line. Again, the first specification of each outcome is a simple differences-in-differences estimation without any control variables, whereas the second one reports the results of the full specification. Here, the full set of control variables are added to control for observable differences and gain statistical precision.

For most specifications and outcomes this table reveals no clear cut evidence for an effect of a drop in people performing community service on hospital production. Let us consider the development of numbers of treated cases first. While it is a common measure of hospital productivity, it can be interpreted in two ways. Whereas an increase could point to a higher productivity of hospitals, i.e. more healthy patients per year, it could also point to higher re-admission rates of patients and has been interpreted in such a way in the previous literature. As can be seen in column (1) and (2) the number of patients treated slightly but insignificantly increases by an additional 214 or 129 cases per year in hospitals experiencing a drop in people performing community service. This equals at most less than 2 percent of the baseline value in the pre-treatment period. According to the common interpretation, this would point to lower treatment success. With this data set, however, there is no possibility to check whether this increment is really caused by higher re-admission rates[35].

To decrease care workload, hospitals might also try to avoid in-patient treatment. If this was the case we should observe a shift towards out-patients. Results in columns (3) to (6) show no significant evidence for this. Whereas the number of out-patients significantly increases when estimated without further controls, it does not seem to change differentially in the two groups once controls are included. The effect of an additional increase of 53 cases reported in column (4) refers to 3 percent of the baseline value[36], but is not significantly different from zero at conventional levels. The additional increment in cases clearly is lower for in-patients. The 75 cases reported in column (6) correspond to 0.7 percent of the pre-treatment period average. Still, the share of in-patients only marginally decreases by 0.001 when estimated using the same specifications. Finally, columns (7) and (8) report hospitals output measured in terms of days of care provided on average per year. Again, results are not significant. On average treated institutions only decreased

[35]Interestingly, there is also a strong time trend towards hospital treatment in general. After the policy change hospitals on average treat around 470 to 820 cases more than in 2009/2010 - even controlling for size. This pattern is also descriptively depicted in the summary statistics in Table 1.1

[36]In 2009/2010 hospitals treated around 1,696 out-patients each year, compare Table 1.1.

Table 1.4: Effects on Hospital Production: Output Quantity

	Case Numbers		Out-patients		In-patients		Days of Care	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Additional Change in Treated Hospitals	214.435	129.041	75.334*	53.384	142.474	75.070	-717.334	-1,044.550
(S.E.)	[226.114]	[107.806]	[42.930]	[32.818]	[197.406]	[92.983]	[1,692.708]	[1,047.624]
Hospital Provider Dummies	No	Yes	No	Yes	No	Yes	No	Yes
Hospital Size Controls	No	Yes	No	Yes	No	Yes	No	Yes
Hospital Patient Mix Controls	No	Yes	No	Yes	No	Yes	No	Yes
Full Set of Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	4,541	4,541	4,541	4,541	4,541	4,541	4,541	4,541
R-squared	0.030	0.674	0.029	0.602	0.030	0.675	0.025	0.685

Notes: The table reports how hospital production changes in response to the suspension of military drafting. The sample is restricted to general hospitals. The dependent variables are the total number of cases in a given hospital and year in column (1)-(2), the number of out-patients in column (3)-(4), the number of in-patients in (5)-(6) and the total number of days of care carried out by a hospital. The key independent variable in all of the specifications reported in this table is an interaction term. It equals one for treated hospitals after the policy change and its coefficient indicates the additional change a "treated" hospital experiences due to the policy change, reported in the first line. The specifications include a variety of hospital controls, regional and patient mix controls as indicated. A number of control variables is included in columns (2), (4), (6) and (8). These are dummy variables for public, private and free or not-for-profit providers, bedclass indicators, summarized in 5 categories (less than 75, 75-124, 125-249, 250-499, 500 or more), controls for the patient composition of hospitals (the share of females, the share of patients aged 75 and above and the share of people that undergo surgery during their hospital stay), dummy variables for the type of hospital (i.e. university hospital, hospital forming part of the German hospital plan, hospital with a hospital provision contract or other hospitals), the share and squared share of apprentices in total staff and regional characteristics. Regional characteristics refers to the degree of regional agglomeration, considering 8 categories ranging from "densely populated urban areas" to "rural county, lightly populated". Including these controls restricts the sample to hospitals with more than 100 beds, because only hospitals with a capacity of 100 beds have to report patient characteristics. All shares are calculated based on staff and patients in the final data set. Standard errors are clustered at the nuts-2 level. *** $p<0.01$, ** $p<0.05$ and * $p<0.1$.

Source: German Hospital and Patient Statistics (2009-2012)

their number of care days by 717 and 1,044 compared to un-treated hospitals. Considering the average duration of stay of less than 9 days per case, this corresponds to 116 patient cases, i.e. around 1 percent of the average case number per hospital. Interestingly, the sign is reversed to the sign of case numbers, indicating a stronger decrease in days performed per patient. As can be seen in Table 1.5, estimating the average duration of stay per case directly, however, does not reveal significant results either (columns (5) and (6)). In summary, some of the results in Table 1.4 are indicative of the fact that health care providers were not fully able to maintain the same output level as before the suspension of drafting. Overall, however, effects are statistically insignificant and economically small. If clinics did not or not strongly adapt their quantitative output levels, did they reduce treatment quality instead? Table 1.5 reports results on patients' health which can be seen as an important indicator of treatment quality. The reported specifications correspond to the ones reported in the previous table. Outcome variables are the number of patients dying in a hospital in a given year in column (1) and (2), the share of patients dying in columns (3) and (4), the average duration of stay per case, column (5) and (6) and the share of patients discharged to another hospital after treatment in column (7) and (8). The binary outcome "death incident" is a particularly reliable indicator of hospital quality because it is robust to different coding behavior across hospitals. The same applies to the duration of stay of a patient or the fact whether he or she was discharged to another hospital, i.e. not fully recovered at the end of the stay in the considered institution. Patterns are rather similar to the results reported in Table 1.4. They do not show any evidence of a loss in treatment quality. Compared to the 250 patients dying in an average hospital per year, the additional dead patient reported in column (2) is not only statistically insignificant but only corresponds to less than 0.5 percent of death incidences. If anything, the negative sign of the share of patients dying reported in the next 2 columns points to a favorable development of death incidences in treated hospitals, but the differential change is also not significantly different from 0 on conventional statistical levels. The same applies to the decreased duration of stay or the likelihood of being discharged to another hospital. A decrease in the duration of stay usually is interpreted as an increase in productivity. A certain health status was obtained earlier[37]. Both estimations report a negative, yet, insignificant coefficient - both statistically as well as economically. The decrease in average duration of stay refers to a 3 percent drop compared to the baseline

[37]Obviously, one could also argue that a shorter duration of stay stands for an increased risk for the patient. People could be discharged too early, leading, for instance, to unsupervised complications or more relapses. As a consequence, one would observe higher re-admission rates. Recall, however, that results on case numbers did not provide any clear cut evidence for more re-admissions (partly, of course, because the data at hand are not feasible for investigating this question). Clearly, the relationship between health status and changes in duration of stay is not linear. Therefore, results from these regressions should be treated with caution.

level, i.e. for more than 33 patients staying for the average duration just 1 is discharged a day earlier. In summary, results point to the fact that hospitals did not react to the policy change in terms of treated cases and objective treatment success.

1.5.3 Measuring Health Outcomes at Patient Level

There are at least two ways to measure patient outcomes when treatment (i.e. the policy change) takes place on a hospital level. As presented in the previous section effects could be estimated on a hospital level. The advantage of this approach is that specifications considering patients' health outcomes and results on other hospital parameters, as, for instance, staffing or costs, are easily comparable as they rest on the exact same estimation strategy. At the same time, estimations of shares of people dying in hospital or average durations of stay include an implicit re-weighting of the results because each hospital, regardless of how many patients are treated in it, is regarded as equally important to the analysis. Policy makers, however, might also be interested in the effects on patients directly. Therefore, an alternative specification considering patient outcomes based on single patient observations is provided. Furthermore, results on the short-run effects on patients which can be identified from within treatment year variation are presented in this paper.

The identification of the effects on patients rests on a symmetric time window around the policy change. To uncover the short-run effects the sample is restricted to 2011. In particular, the health outcomes of patients hospitalized six months before and after the cut-off date are compared. A key concern with this simple difference is that the health outcomes of patients might systematically vary on the two sides of the cut-off date, i.e. patients hospitalized between January and June might a priori have different average durations of stay and mortality levels than patients hospitalized between July and December, for instance, due to seasonality in diseases. In order to avoid biased estimates, I use patients from the previous year as an additional control group. This control group is then comprised of patients hospitalized (and discharged) between January, 1st 2010 and December 31, 2010. All of these patients where unaffected by the suspension of military drafting and no other policy change occurred at the control cut-off date. The regarding regression model can be written in the following way:

$$y_{ihm} = \beta_0 + \beta_1 above_{ih} + \beta_2 after_{im} + \beta_3 (above_{ih} \cdot after_{im}) + \sum_m \theta_m D_{im} + \lambda' X_{ihm} + \varepsilon_{ihm} \tag{1.3}$$

where $above_{ih}$ is a binary variable equal to one if a patient i was treated in a hospital h that was highly affected by the policy change, i.e. it switches on for patients in hospitals whose

Table 1.5: Effects on Hospital Production: Patient Outcomes

	No. of Patients Dying		Share of Patients Dying		Staying Duration		Discharged to Hospital	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Additional Change in Treated Hospitals	0.282	1.140	-0.001	-0.000	-0.390	-0.296	-0.003	-0.001
(S.E.)	[5.433]	[3.756]	[0.001]	[0.000]	[0.234]	[0.211]	[0.002]	[0.002]
Hospital Provider Dummies	No	Yes	No	Yes	No	Yes	No	Yes
Hospital Size Controls	No	Yes	No	Yes	No	Yes	No	Yes
Hospital Patient Mix Controls	No	Yes	No	Yes	No	Yes	No	Yes
Full Set of Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	4,541	4,541	4,541	4,541	4,541	4,541	4,541	4,541
R-squared	0.024	0.600	0.003	0.470	0.008	0.282	0.005	0.232

Notes: The table reports how hospital production in terms of patient outcomes changes in response to the suspension of military drafting. The sample is restricted to general hospitals. The dependent variables are the average numbers of patients dying while hospitalized per year and hospital in column (1)-(2), share of patients dying in a given hospital (3)-(4), the average staying duration in a given hospital (5)-(6) and the share of patients discharged to another hospital (7)-(8). The key independent variable in all of the specifications reported in this table is an interaction term. It equals one for treated hospitals after the policy change and its coefficient indicates the additional change a "treated" hospital experiences due to the policy change. The effect is reported in the first line. The specifications include a variety of hospital controls, regional and patient mix controls as indicated. The full specification includes dummy variables for public, private and free or not-for-profit providers, bedclass indicators summarized in 5 categories (less than 75, 75-124, 125-249, 250-499, 500 or more), controls for the patient composition of hospitals (share of females, the share of patients aged 75 and above and the share of people that undergo surgery during their hospital stay), dummy variables for the type of hospital (i.e. university hospital, hospital forming part of the German hospital plan, hospital with a hospital provision contract or other hospitals), the share and squared share of apprentices in total staff and regional characteristics. Regional characteristics refers to the degree of regional agglomeration, considering 8 categories ranging from "densely populated urban areas" to "rural county, lightly populated". Including these controls restricts the sample to hospitals with more than 100 beds, because only hospitals with a capacity of 100 beds have to report patient characteristics. All shares are calculated based on staff and patients in the final data set. Standard errors are clustered at the nuts-2 level. *** $p<0.01$, ** $p<0.05$ and * $p<0.1$.

Source: German Hospital and Patient Statistics (2009-2012)

share of people performing community service was above zero before 2011. The coefficient β_1 then captures the differences between highly affected and unaffected hospitals. $after_{im}$ is a binary indicator equal to one if patient i was hospitalized in the second half of a given year, i.e. between months m July and December. β_2 captures differences in-between patients treated in different parts of the year. The interaction between these previous two variables ($above_{ih} \cdot after_{im}$), switches on for patients hospitalized in the second half of 2011, i.e. after the suspension of military drafting. The coefficient β_3 then is the differences-in-differences effect of interest. To allow for additional heterogeneity in seasonal effects that may arise because patients are affected by different conditions over the year, I control for a full set of month-of-hospitalization dummies. X_{ihm} is a vector of additional control variables including the same variables as in the previous specifications and indicated in the regarding tables. Equation (1.3) is estimated separately for a patient's duration of stay and likelihood of dying. To account for correlation of patient outcomes within regions, standard errors ε_{ihm} are clustered at the nuts-2 level[38].

Results for the probability of dying in a hospital and the average duration of stay on patient level are presented in Tables 1.6 and 1.7. Recall that when average numbers or share of patients dying in hospital where estimated on hospital level, no effect on either of the two indicators was detectable. Interestingly, when estimated on patient level small effects on the probability of dying in a hospital can be observed. As can be seen in columns (1) to (4) of Table 1.6 the effect is smaller than 0.0 percentage points unless all control variables are included (column (5)). The full specification indicates a small positive effect on the likelihood of dying 0.1 percentage points which stands for 4 percent of the baseline value of 2.24 percentage points. Considering the decrease in the likelihood of dying over time[39], the effect of the policy change almost fully compensates the favorable development in death probability all hospitals experience. This means that despite controlling for hospital size, hospital averages cover up a small effect on the risk of dying for all patients. Note, however, that the R-squared of these estimations is also really low even despite the fact that a full set of month of hospitalization controls is included. One could also suspect the significance of these results to be driven by the high number of patient observations. Whereas the estimations based on hospital averages were relying on around 4,500 observations, the sample here includes more than 28 million observations. Table 1.7, presenting results on patients durations of stay, indicates that this is not the case. Based on the same number of observations results on the length of stay are not significantly different from 0. Hence, they deliver the same result as the

[38] Clustering standard errors on the hospital level to account for within hospital variation instead only slightly increases significance for the effect on duration of stay in one specification. All other tests deliver the exact same significance levels. Results are available on request.

[39] Compare the summary statistics presented in Table 1.1 which indicate a drop of 0.13 percentage points.

previous specifications. The policy change did not have any effect on the average duration of stay. Whereas results on hospital level lead to an insignificant coefficient of between -0.29 and -0.39 days, results on the patient level are even closer to zero (between -0.04 and -0.06 days).

Table 1.6: Death Incidents on Patient Level

	(1)	(2)	(3)	(4)	(5)
Additional Change in Treated Hospitals	0.000*	0.000*	0.000*	0.000**	0.001**
(S.E.)	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Additional Staff for Free/Not-for-Profit Providers		0.002**			0.001
(Baseline Public Hospitals)		[0.001]			[0.000]
Additional Staff in Private Hospitals		0.001			-0.001
		[0.001]			[0.000]
Additional Staff in Hospitals with 75-124 Beds			0.010***		0.004*
(Baseline <75 Beds in Hospital)			[0.003]		[0.002]
Additional Staff in Hospitals with 125-249 Beds			0.012***		0.006***
			[0.003]		[0.002]
Additional Staff in Hospitals with 250-499 Beds			0.013***		0.008***
			[0.003]		[0.002]
Additional Staff in Hospitals with 500 and more			0.012***		0.009***
			[0.003]		[0.002]
Share of People Obtaining Surgery				-0.010***	-0.009***
				[0.002]	[0.001]
Share of Female Patients				-0.018***	-0.007
				[0.004]	[0.004]
Share of Patients 75 and Above				0.074***	0.085***
				[0.005]	[0.005]
Hospital Provider Dummies	No	Yes	No	No	Yes
Hospital Size Controls	No	No	Yes	No	Yes
Hospital Patient Mix Controls	No	No	No	Yes	Yes
Full Set of Controls	No	No	No	No	Yes
Observations	28,135,721	28,135,721	28,135,721	28,135,721	28,135,721
R-squared	0.000	0.000	0.000	0.001	0.001

Notes: The table reports how the likelihood of dying in hospital changes in response to the suspension of military drafting. The sample is restricted to patients in general hospitals. The dependent variable is a binary indicator for death. The key independent variable in all of the specifications reported in this table is an interaction term. It equals one for treated hospitals after the policy change and its coefficient indicates the additional change a "treated" hospital experiences due to the policy change. The first line reports the additional change in the likelihood of dying in treated hospitals due to the policy change (the differences-in-differences estimator). The specifications include a variety of hospital, regional and patient mix controls as indicated. Dummy variables for public, private and free or not-for-profit providers are included in column (2). Bedclass indicators, summarized in 5 categories (less than 75, 75-124, 125-249, 250-499, 500 or more), are included in column (3). The specification in column (4) controls for the patient composition of hospitals by including the share of females, the share of patients aged 75 and above and the share of people that undergo surgery during their hospital stay as control variables. Additionally to all of the above, the specification including the full set of control variables, presented in column (5), includes dummy variables for the type of hospital (i.e. university hospital, hospital forming part of the German hospital plan, hospital with a hospital provision contract or other hospitals), the share and squared share of apprentices in total staff and regional characteristics. Regional characteristics refer to the degree of regional agglomeration, considering 8 categories ranging from "densely populated urban areas" to "rural county, lightly populated". Note that only hospitals with a capacity of 100 beds have to report patient characteristics. Standard errors are clustered at the nuts-2 level. *** $p < 0.01$, ** $p < 0.05$ and * $p < 0.1$.

Source: German Hospital and Patient Statistics (2010-2011)

Table 1.7: Duration of Stay on Patient Level

	(1)	(2)	(3)	(4)	(5)
Additional Change in Treated Hospitals	-0.043	-0.044	-0.040	-0.061	-0.046
(S.E.)	[0.044]	[0.045]	[0.043]	[0.053]	[0.052]
Additional Staff for Free/Not-for-Profit Providers		-0.190			-0.165
(Baseline Public Hospitals)		[0.168]			[0.137]
Additional Staff in Private Hospitals		-0.208			-0.017
		[0.177]			[0.157]
Additional Staff in Hospitals with 75-124 Beds			-0.381		-0.147
(Baseline <75 Beds in Hospital)			[0.785]		[0.800]
Additional Staff in Hospitals with 125-249 Beds			-0.754		-0.351
			[0.687]		[0.719]
Additional Staff in Hospitals with 250-499 Beds			-0.459		-0.213
			[0.710]		[0.747]
Additional Staff in Hospitals with 500 and more			-0.114		0.018
			[0.707]		[0.746]
Share of People Obtaining Surgery				-2.469***	-2.450***
				[0.417]	[0.414]
Share of Female Patients				-5.466***	-4.725***
				[0.912]	[0.894]
Share of Patients 75 and Above				-1.562	0.183
				[1.579]	[1.840]
Hospital Provider Dummies	No	Yes	No	No	Yes
Hospital Size Controls	No	No	Yes	No	Yes
Hospital Patient Mix Controls	No	No	No	Yes	Yes
Full Set of Controls	No	No	No	No	Yes
Observations	28,135,721	28,135,721	28,135,721	28,135,721	28,135,721
R-squared	0.006	0.006	0.007	0.008	0.010

Notes: The table reports how patients' staying duration changes in response to the suspension of military drafting. The sample is restricted to patients in general hospitals. The dependent variable is the number of days a patient stayed in a given hospital. The key independent variable in all of the specifications reported in this table is an interaction term. It equals one for treated hospitals after the policy change and its coefficient indicates the additional change a "treated" hospital experiences due to the policy change. The first line reports the additional change in staying duration in treated hospitals due to the policy change (the differences-in-differences estimator). The specifications include a variety of hospital, regional and patient mix controls as indicated. Dummy variables for public, private and free or not-for-profit providers are included in column (2). Bedclass indicators, summarized in 5 categories (less than 75, 75-124, 125-249, 250-499, 500 or more), are included in column (3). The specification in column (4) controls for the patient composition of hospitals by including the share of females, the share of patients aged 75 and above and the share of people that undergo surgery during their hospital stay as control variables. Additionally to all of the above, the specification including the full set of control variables, presented in column (5), includes dummy variables for the type of hospital (i.e. university hospital, hospital forming part of the German hospital plan, hospital with a hospital provision contract or other hospitals), the share and squared share of apprentices in total staff and regional characteristics. Regional characteristics refer to the degree of regional agglomeration, considering 8 categories ranging from "densely populated urban areas" to "rural county, lightly populated". Note that only hospitals with a capacity of 100 beds have to report patient characteristics. Standard errors are clustered at the nuts-2 level. *** p<0.01, ** p<0.05 and * p<0.1.

Source: German Hospital and Patient Statistics (2010-2011)

1.5.4 The Drop in People Performing Community Service and Hospital Staffing

The previous section has shown that hospitals were largely able to maintain their output levels in terms of quantity as well as quality. But then, how did they compensate the loss of draftees? If they did not adjust output levels did they hire other staff? As the majority of draftees in hospitals performed tasks in care support, nurses can be seen as the closest substitute to people performing community service and the staff group most likely to compensate the work of draftees[40]. Table 1.8 provides results on how hospital's labor input from nurses changed through the suspension of drafting. I apply the same specification as above, with sets of control variables varying as indicated in the table. Again, the first column reports the pure difference in development over time between the treatment and control group without including controls. To shed light on the changes in staffing several different staffing measures are constructed. I first consider overall labor input by nurses by estimating yearly average FTEs. This accounts for different types of employment relationships. It reveals that hospitals that experienced a drop in people performing community service did not hire more additional nurses than hospitals that did not experience a drop. The coefficient is positive in specifications without control variables or when controlling for the full set of controls, but is not significantly different from zero in any of the presented or further tested specifications.

It is, of course, possible that hospitals implicitly adapted their labor input by changing the workload the same personnel had to manage. To examine the complete picture of care staffs' labor input, 3 more indicators of labor input are reported in Table 1.8. The effects on nurses per bed are reported in columns (3) and (4). In both specifications presented, as well as in all other specifications including the control variables discussed above, the differences-in-differences coefficient is insignificant and rather close to zero, indicating that the capacity nurses had to cover did not change differentially in the two groups. Obviously, the pure capacity of the hospital does not always equate the actual workload per person, therefore, another indicator considering the workload of nurses in terms of care days is constructed[41]. The baseline and main specifications are reported in columns (5) and (6). Again, the differences-in-differences coefficients show no response in reaction to the suspension of drafting. Finally, the same applies to the workload of nurses in terms of patient case numbers. The workload in terms of case numbers reflects a different aspect of workload because some labor intensive tasks, as, for instance, paper

[40] Note that the group "nurses" subsumes trained nurses as well as untrained but fully paid employees in care, but excludes specialized care personnel, such as radiographers or physiotherapists and volunteers. In Germany nursing is a 3 year vocational training.

[41] It reflects how many hours of care (i.e. invoiced care days x 24) were actually performed per working hours of a FTE equivalent (i.e. per FTE x 220 [working days per year] x 8 [hours])

work or general examinations, only have to be performed in the context of an intake or a discharge of a person. Regressions reported in column (7) and (8) show that the number of patient cases in a hospital did not develop differently in the two groups, revealing that hospitals also did not react in this dimension of care staff's workload either. When all hospital characteristics are controlled for, the coefficient raises to 1.6, indicating that treated hospitals experienced an additional increment of patient cases per nurse, but considering the standard error of 1.1 this coefficient is not significantly different from zero at conventional levels. It is, however, important to note that none of the measures referring to FTEs takes extra hours into account.

If hospitals did not adapt in terms of trained and untrained nurses, did they compensate the labor input in terms of other staff groups instead? To shed light on this question Table 1.9 reports the results for specialized medical care staff and apprentices, two other staff groups, that could be regarded as similar to draftees. Hospitals could educate more apprentices for two reasons. Firstly, among all staff groups in a hospital apprentices are most likely to carry out low-skilled and assisting tasks, in particular in care, whereas their salaries are comparably low. Therefore, hiring more apprentices could be a cost conscious way of replacing people performing community service. Secondly, it could be possible that there is not enough trained staff available on the labor market. Training apprentices could then be the only option to replace care staff in the long-run. Results in column (1) and (2) show that this was not the case. No matter which control variables included, the differences-in-differences estimator is never distinguishable from zero. Another short-run alternative could be hiring more specialized medical care staff, for instance, dietary advisors or radiographers. These are usually more costly if, however, there is not enough other care staff available they could be employed instead. Column (3) and (4) show that this was not the case. Without controlling for hospital characteristics, there seems to be an additional increase in hospitals experiencing a decrease in people performing community service, but the coefficient moves a lot closer to zero and becomes insignificant once hospital and regional characteristics are controlled for. In summary, no other staff group performing care or care related tasks was adjusted measurably in response to the drop out of people performing community service.

Similar patterns hold true for the effects on doctoral staff presented in Table 1.10. The first two columns present results on numbers on doctors regardless of type of contract. This measure is added because physicians are more likely to work hours over-time than the other staff groups discussed above. Therefore, hiring more (part-time) physicians could help hospitals in increasing labor input without necessarily adapting the amount of contracted hours. Column (3) and (4) present the effect on full-time equivalences, whereas columns (5) and (6) consider the workload in terms of patient hours relative to

Table 1.8: Effects on Measures of Staffing: Nurses

	FTE of Care Staff		Nurses per Bed		Workload Care Hours		Workload Cases	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Additional Change in Treated Hospitals	-4.006	-1.371	-0.003	0.000	0.056	0.034	1.761	1.674
(S.E.)	[5.259]	[2.987]	[0.010]	[0.009]	[0.081]	[0.094]	[1.232]	[1.157]
Additional Staff for Free/Not-for-Profit Providers		40.300***		0.009		-0.440**		-1.420
(Baseline Public Hospitals)		[12.072]		[0.014]		[0.164]		[1.573]
Additional Staff in Private Hospitals		-6.204		0.010		-0.478**		-5.394***
		[7.542]		[0.016]		[0.207]		[1.382]
Additional Staff in Hospitals with 75-124 Beds		26.025***		-0.091**		-0.528		-7.534
(Baseline <75 Beds in Hospital)		[7.591]		[0.043]		[0.452]		[6.992]
Additional Staff in Hospitals with 125-249 Beds		68.923***		-0.105**		-0.377		-4.969
		[6.052]		[0.045]		[0.401]		[6.451]
Additional Staff in Hospitals with 250-499 Beds		159.015***		-0.100**		-0.423		-8.041
		[8.361]		[0.047]		[0.407]		[6.463]
Additional Staff in Hospitals with 500 and more		408.539***		-0.084		-0.557		-11.538*
		[17.894]		[0.052]		[0.398]		[6.733]
Share of People Obtaining Surgery		-13.101		-0.098*		-0.593		35.634***
		[12.193]		[0.050]		[0.597]		[5.111]
Share of Female Patients		-201.342***		-0.899***		7.066***		101.923***
		[37.077]		[0.121]		[1.368]		[10.095]
Share of Patients 75 and Above		-99.077***		-0.016		-0.297		-26.103***
		[22.481]		[0.088]		[1.187]		[7.056]
Hospital Controls	No	Yes	No	Yes	No	Yes	No	Yes
Regional Controls	No	Yes	No	Yes	No	Yes	No	Yes
Patient Mix Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	4,541	4,541	4,541	4,541	4,541	4,541	4,541	4,541
R-squared	0.022	0.718	0.003	0.225	0.005	0.144	0.001	0.214

Notes: The table reports how staffing with care personnel changes in response to the suspension of military drafting. The dependent variables measure the number of FTE per installed hospital bed in (1)-(2), how many care hours (reported care days x 24) were performed per actual working hours of a yearly average full-time equivalent (FTE x 220 (working days per year) x 8 hours) in column (3)-(4), the case numbers per hospital-year relative to the number of care staff full-time equivalences in columns (5) - (6) (includes trained as well as untrained personnel working in care) and the workload in terms of treated patient cases per yearly average FTE in columns (7) - (8). The first line reports the additional change in staffing in treated hospitals due to the policy change (the differences-in-differences estimator). It is the coefficient of an interaction term switching on for treated hospitals after the policy change. A hospital is considered treated, if it employed people performing community service before the supension of drafting. The after period is equal to 2012. The sample includes all general hospitals in Germany with non-missing values for all variables (constant sample). This excludes very small hospitals with less than 100 beds. The first specification always shows the pure difference-in-difference estimator and the second specification includes the full set of control variables to gain precision. The full set of control variables includes dummy variables for public, private and free or not-for-profit providers, bedclass indicators, summarized in 5 categories (less than 75, 75-124, 125-249, 250-499, 500 or more), controls for the patient composition of hospitals (i.e. the share of females, the share of patients aged 75 and above and the share of people that undergo surgery during their hospital stay), dummy variables for the type of hospital (i.e. university hospital, hospital forming part of the German hospital plan, hospital with a hospital provision contract or other hospitals), the share and squared share of apprentices in total staff and regional characteristics (considering 8 categories ranging from "densely populated urban areas" to "rural county, lightly populated"). Standard errors are clustered at the nuts-2 level. *** p<0.01, ** p<0.05 and * p<0.1.

Source: German Hospital and Patient Statistics (2009-2012)

Table 1.9: Effects on Measures of Staffing: Other Medical Staff and Apprentices

	Apprentices		Other Medical Staff	
	(1)	(2)	(3)	(4)
Additional Change in Treated Hospitals	0.241	-0.702	-0.625	1.140
(S.E.)	[1.444]	[1.051]	[6.683]	[3.284]
Hospital Controls	No	Yes	No	Yes
Regional Controls	No	Yes	No	Yes
Patient Mix Controls	No	Yes	No	Yes
Observations	4,541	4,541	4,541	4,541
R-squared	0.024	0.575	0.016	0.766

Notes: The table shows how the staffing with apprentices (columns (1)-(2)) and other medical staff (columns (3)-(4)) respond to the suspension of military drafting. Other medical staff refers to specialised and trained care staff, including for instance physiotherapists, radiographers or dietary advisors but excluding nurses. Both groups are given in yearly average FTE. The first specification always shows the pure difference-in-difference estimator and the second specification includes the full set of control variables to gain precision. The full set of control variables includes dummy variables for public, private and free or not-for-profit providers, bedclass indicators, summarized in 5 categories (less than 75, 75-124, 125-249, 250-499, 500 or more), controls for the patient composition of hospitals (i.e. the share of females, the share of patients aged 75 and above and the share of people that undergo surgery during their hospital stay), dummy variables for the type of hospital (i.e. university hospital, hospital forming part of the German hospital plan, hospital with a hospital provision contract or other hospitals), the share and squared share of apprentices in total staff and regional characteristics (considering 8 categories ranging from "densely populated urban areas" to "rural county, lightly populated"). Shares are calculated based on staff and patients in the final data set. The sample is restricted to general hospitals with more than 100 beds. Standard errors are clustered at the nuts-2 level. *** p<0.01, ** p<0.05 and * p<0.1.

Source: German Hospital and Patient Statistics (2009-2012)

working hours of doctors (parallel to the measure for nurses presented above) the last two columns present the labor input relative to patient case numbers. Eventually, none of the dimensions indicates a differential change in treatment and control group hospitals.
In conclusion, hospitals did not notably adapt any of the contracted labor input related to health production and patient care in response to the suspension of drafting.

1.5.5 The Reduction in Low-Skilled Personnel and Hospital Costs

So far we have established that hospitals were largely able to maintain their output levels while not adapting any of their contracted labor input. According to the hospital production function discussed above, we then should observe higher spending on capital input indicating that hospitals moved along the hospital production niveau curve, i.e. produce the same output with a different ratio of capital to labor. Do the empirical findings corroborate these theoretical predictions?
Table 1.11 presents results on hospital costs showing that empirical findings are indeed in line with theoretical predictions. As a first overview, columns (1) and (2) show overall hospital expenditure. This is a summary measure for all costs of the hospital that are not covered by public authorities, including material costs, staff costs, training costs just as well as costs for hospital administration. Overall hospital costs increased by additional

Table 1.10: Effects on Measures of Staffing: Doctoral Staff

	Numbers of Doctors		FTE of Doctors		Workload Care Hours		Workload Cases	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Additional Change in Treated Hospitals	0.384	4.413	-0.523	0.841	-0.460	-0.318	-4.039	-4.456
(S.E.)	[9.005]	[5.121]	[3.098]	[1.595]	[0.614]	[0.578]	[6.224]	[7.112]
Hospital Controls	No	Yes	No	Yes	No	Yes	No	Yes
Regional Controls	No	Yes	No	Yes	No	Yes	No	Yes
Patient Mix Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	4,541	4,541	4,541	4,541	4,541	4,541	4,541	4,541
R-squared	0.020	0.753	0.018	0.755	0.004	0.070	0.001	0.059

Notes: The tables shows that contracted doctoral staffing does not change in response to the suspension of military drafting. The dependent variables are numbers of doctors in columns (1)-(2), yearly average full-time equivalences of doctors in columns (3)-(4), workload in terms of performed care hours per working hours of a yearly average full-time equivalent (5)-(6) and numbers of treated cases per yearly average full-time equivalent in column (7)-(8). Thereby, care hours refers to the number of provided care days x 24 and the actual working hours are calculated as follows: FTE x 220 (working days per year) x 8 hours. All types of doctors are included.The first specification always shows the pure difference-in-difference estimator and the second specification includes the full set of control variables to gain precision. The full set of control variables includes dummy variables for public, private and free or not-for-profit providers, bedclass indicators, summarized in 5 categories (less than 75, 75-124, 125-249, 250-499, 500 or more), controls for the patient composition of hospitals (i.e. the share of females, the share of patients aged 75 and above and the share of people that undergo surgery during their hospital stay), dummy variables for the type of hospital (i.e. university hospital, hospital forming part of the German hospital plan, hospital with a hospital provision contract or other hospitals), the share and squared share of apprentices in total staff and regional characteristics (considering 8 categories ranging from "densely populated urban areas" to "rural county, lightly populated"). Shares are calculated based on staff and patients in the final data set. The sample is restricted to general hospitals with more than 100 beds. Standard errors are clustered at the nuts-2 level. *** p<0.01, ** p<0.05 and * p<0.1.

Source: German Hospital and Patient Statistics (2009-2012)

2.4 million Euros in hospitals that experienced a drop in people performing community service. Considering that hospitals on average spent 46.7 million Euros per year in the pre-policy period and 52.1 million in the after policy period this corresponds to around 5 percent of the average expenditure of hospitals or an increase to the general time trend of 44 percent.

Considering the single cost components in more detail, as presented in columns (3) to (10), leads to the insight that the largest share of this increment is caused by material costs (45 percent) and expenditure on doctoral staff (24 percent). Material costs include various types of costs related to the treatment of the patient, in particular, medical consumables like drugs, bandages, instruments, therapeutic appliances, blood and plasma and can be seen as a way to measure capital input of the hospital. An increment in average hospital material expenditure indicates that more capital was needed to treat, as shown before, a rather stable number of patients. This not only points to pure economic inefficiencies in the production process, but can also have potentially harmful consequences for patients not measured by the objective output indicators presented above. Cawley et al. (2006), for instance, show that more capital input can stand for greater use of psychoactive drugs, which does not necessarily lead to a higher likelihood to die but at the same time might not be the ideal way to treat patients. This evidence is in line with studies finding that less skilled doctors (i.e. lower labor input on the intensive margin) cause higher treatment costs (Doyle et al. 2010).

Interestingly, splitting up total hospital expenditure also reveals an increment of expenditure on doctoral staff. On average, as reported in columns (9) and (10) of Table 1.11, affected hospitals spent slightly more than half a million Euros on physicians extra, while, as shown above, not hiring significantly more doctors. An average hospital employed 88 doctors after the policy change, i.e. this effect equals 565 additional Euros per doctor each month after losing the people performing community service. With this data set it is not possible to distinguish whether physicians earned more per contracted hour or worked more hours than contracted[42], but it is very clear that at least one of the two must be true. Given that hospitals paid around 57 Euros per contracted hour of labor input on doctors before, 565 additional Euros per month stand for 2.5 extra hours of work each week. Assuming that (only) hours worked overtime increased, these would add up to 880 hours per year in total[43]. Volunteers replacing people performing community service offered around 1100 hours. Interestingly, this would mean that the additional labor input by doctors and volunteers add up to around the loss of hours performed by people performing community service (1840 hours). As shown in columns (5) to (8) of Table 1.11,

[42] Recall that only contracted hours are reported, whether staff works over time is not documented.

[43] If hospitals paid higher salaries the additional 565 Euros would stand for an increase of around 3.50 Euros per hour.

Table 1.11: Effects on Hospital Expenditure

	Total Exp.		Material Costs		Exp. Care Staff		Exp. Apprentices		Exp.Doctoral Staff	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Additional Change in Treated Hospitals	2.720*	2.401*	1.149*	1.088**	0.237	0.182	0.022	0.012	0.650**	0.597***
(S.E.)	[1.479]	[1.218]	[0.600]	[0.537]	[0.228]	[0.180]	[0.020]	[0.016]	[0.286]	[0.207]
Hospital Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Regional Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Patient Mix Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	4,535	4,535	4,535	4,535	4,535	4,535	4,535	4,535	4,535	4,535
R-squared	0.019	0.754	0.017	0.740	0.024	0.708	0.013	0.479	0.022	0.744

Notes: The table shows how hospital expenditure changes in response to the suspension of military drafting. The dependent variables are total hospital expenditure per hospital and year in Euros in columns (1)-(2), yearly material costs in columns (3)-(4), yearly expenditure on care staff in columns (5)-(6), overall training costs, including expenditure on training apprentices in columns (7) and (8) as well as yearly expenditure on doctoral staff in columns (9) - (10). The key independent variable in all specifications is an interaction term switching on for treated hospitals after the policy change. A hospital is considered treated if it employed people performing community service before the supension of drafting. The after period is equal to 2012. The sample includes all general hospitals in Germany with non-missing values for all variables (constant sample). This excludes very small hospitals with less than 100 beds. The first specification always shows the pure difference-in-difference estimator and the second specification includes the full set of control variables to gain precision. The full set of control variables includes dummy variables for public, private and free or not-for-profit providers, bedclass indicators, summarized in 5 categories (less than 75, 75-124, 125-249, 250-499, 500 or more), controls for the patient composition of hospitals (i.e. the share of females, the share of patients aged 75 and above and the share of people that undergo surgery during their hospital stay), dummy variables for the type of hospital (i.e. university hospital, hospital forming part of the German hospital plan, hospital with a hospital provision contract or other hospitals), the share and squared share of apprentices in total staff and regional characteristics (considering 8 categories ranging from "densely populated urban areas" to "rural county, lightly populated"). Standard errors are clustered at the nuts-2 level. *** p<0.01, ** p<0.05 and * p<0.1.

Source: German Hospital and Patient Statistics (2009-2012)

there is no additional spending on care staff and apprentices.
In summary, the findings on hospital expenditure show two things. Firstly, the drop in low-skilled labor leads to a higher capital to labor input ratio. While hospitals are able to maintain their output levels in terms of quantity and measurable quality this still might have potentially harmful consequences for patients. Secondly, while compensating the loss in labor input by utilizing more capital, hospitals also spend more money on doctoral labor, which is complementary to capital input. Interestingly, the additional spending amounts to around 880 hours equivalences of labor, which in sum with hours worked by volunteers equals approximately the average decrease in labor input by people performing community service.

1.6 Conclusion

This paper investigates labor in hospital production. It sheds light on how hospital production, staffing, and costs react to a sudden reduction of labor input from low-skilled staff. Providing evidence on this requires very comprehensive data on hospital's in- as well as outputs. Therefore, this paper uses administrative data from the German Hospital and Patient Statistics, covering all of the hospitals and clinics of the country including the vast majority of their patients. The identification strategy of this paper exploits quasi-experimental variation in staffing induced by the indefinite suspension of military drafting in Germany in 2011.
This policy change, in combination with newly available rich administrative data, allows to add to the existing literature by providing one of the first papers to use exogenous variation, which affects hospital staffing, but does not originate in the health system. Additionally, the indefinite suspension of drafting enables us to study a much longer time horizon than most other studies, as previous sources of variation (strikes or seasonal fluctuations) are typically short-lived. By exploiting exogenous variation from the suspension of mandatory military service this paper also contributes to the understanding of the economic effects of drafting and its substitute services.
Empirical results provide several interesting insights. Firstly, hospitals impacted by the policy change neither decrease their quantitative output nor their measurable output quality.
Secondly, medical institutions also do not alter their staffing structure in response to the suspension of drafting. There is no evidence on a differential change in number of full-time equivalences of nurses, specialized care staff, apprentices or doctors between hospitals affected and unaffected by the policy change. The same applies to various other measures of staff's workload, like, for instance, the number of cases or beds per nurse and doctor

or the number of care hours performed.
Thirdly, considering hospital's costs reveals interesting patterns on capital input as well as on labor expenditure. For one thing, the drop in low-skilled labor leads to a higher capital to labor input ratio. While hospitals are able to maintain their output levels in terms of quantity and measurable quality this still might have potentially harmful consequences for patients. For another thing, while compensating the loss in labor input by utilizing more capital, hospitals also spend more money on doctoral labor, either to finance overtime or on higher wages. Interestingly, the additional spending is equivalent to around 880 hours of labor from doctors, which in sum with hours worked by volunteers equals approximately the average decrease in labor input by people performing community service. There is no additional spending on care staff and apprentices.

1.A Appendix

1.A.1 German Hospital and Patient Statistics (2009-2012)

The administrative data in the German Hospital and Patient Statistics is an important informational source for health policy and serves as the planning basis for the federal states involved in financing of hospitals and clinics. Every hospital and clinic is obliged by law to report detailed information on a variety of figures to the Federal Ministry of Health. The data are collected and administered by the statistical offices of the German federal states. General conditions regarding these statistics are regulated in §28 of the KHG. To focus on the periods around the policy change, I restrict the Hospital and Patient Statistics to the years 2009 to 2012. The basic sample consists of between 1571 (2009) and 1495 (2012) general hospitals reporting detailed figures on hospital production, staff numbers, hospital costs and a variety of other measures that serve as control variables.

People performing community service: The identification strategy of this paper rests on the number of people performing community service in the year 2010, the year prior to the policy change. Treatment and control group in all specifications are distinguished by the extent to which hospitals rely on labor input by people performing community service. The main specification compares hospitals that did not report any people performing community service in 2010 to those that did. Hospitals are not only obligated by law to report staff, they also have clear incentives to report all people performing community service, because their salaries were partly paid by public authorities. Therefore, I also include hospitals that reported a missing for people performing community service in the control group, assuming they simply did not report zeros.

Hospital output variables: To analyze how hospitals' output reacts to the policy change, I construct several measures on numbers of patients treated and treatment duration. *Case numbers* is an overall head count of all in- and outpatients treated in a hospital in a given year. *Number of out-patients* refers to numbers of registered cases for which hospitals reported a duration of stay of a day or less in general hospitals. This is done because registering a patient with a duration of stay of one day is very common even if the actual hospital stay takes less than 24 hours and does not include staying over night. *In-patients* are identified by a duration of stay of two days and above. An analysis of the development and interaction of these types of cases allows us to identify one important potential way of reacting to the policy change if original staffing levels cannot be retained. *Days of care* is a summary measure of the total number of care days performed in a hospital in a given year regardless of case numbers and severity.

Further output measures refer to patients health as objectively as possible. *Mortality* is measured in absolut numbers of in-hospital death incedences as well as in share of people dying in hospital (=case fatality rate). Mortality is the most severe clinical outcome, but is also commonly used as a quality indicator for hospitals. One clear advantage of this outcome is, that it is robust to different coding behavior of hospitals (Wissenschaftliches Institut der AOK 2007: p.29) and thereby easily comparable across hospitals. This can be of relevance because hospitals record their own data and know that this data can be used for quality evaluation. Therefore, despite the fact that there are strong incentives to report data correctly as to being paid accordingly, there might be an incentive to down play adverse events like complications. *Average duration of stay* indicates the average du-

ration of stay of all in-patients in the final sample in a hospital. Hospitals have a strong incentive to discharge patients as soon as possible, i.e. as soon as a patient reaches the necessary level of health to be discharged. A longer duration of stay means that hospitals need more time to reach that level, indicating adverse effects on patients' health. *Discharges to other hospitals* provides the share of patients that at the end of their stay in a clinic cannot be discharged but need to be transferred to another hospital. The reason for this transferral cannot be identified in the data, therefore, it is unclear whether a patient has not yet recovered, is eligible for a rehabilitation clinic or the hospital does not have the capacity to treat him or her any further.

Labor input variables: The main dependent variables to measure labor input are the numbers (regardless of type or scope of employment relationship) and yearly average full-time equivalences of all low, medium and high-skilled staff of hospitals. Staff head count refers to the number of all directly employed people at any institution as matters stand on December, 31 each year. This does not account for different types of employment relationships. Therefore, information on yearly average FTE is included. FTEs are calculated based on the working hours of all contracts over the course of the year, for instance, a full-time employed nurse working from January until March, would be reported as 0.25 FTE. The same applies to most other groups considered, except apprentices who are converted with a factor of 9.5 to 1 or 6 to 1, depending on their type.

To reflect various aspects of care staffs' labor input, I define 3 more measures of labor input by doctoral and care staff. *Nurses/doctors per bed* refers to yearly average full-time equivalences per installed hospital bed. *Workload of nurses and doctors in terms of care hours* is an indicator of how many care hours were performed per actual working hours of a yearly average full-time equivalent[44]. *Workload of nurses/doctors in terms of case numbers* relates the number of cases per hospital and year to the number of full-time equivalences.

Cost measures: The main dependent variables to measure hospitals costs and expenditure on a variety of assets are hospital's overall expenditure and expenditure split up by single staff groups and material in Euros per year. Overall hospital costs is a summary measure for all costs of the hospital that are not covered by public authorities, including material costs, staff costs, training costs just as well as costs for hospital administration. Costs for the single staff groups are split up in eleven subcategories in the original data set (including for instance doctoral services, care services, medical-technical services and administration). This paper considers doctoral services, cares services and expenditure on staff training in more detail. All of these are measured in yearly sum per hospital grouped by staff category. Material costs include various types of costs related to the treatment of the patient. In particular, medical consumables like drugs, bandages, instruments, therapeutic appliances, blood and plasma are contained in this summary measure.

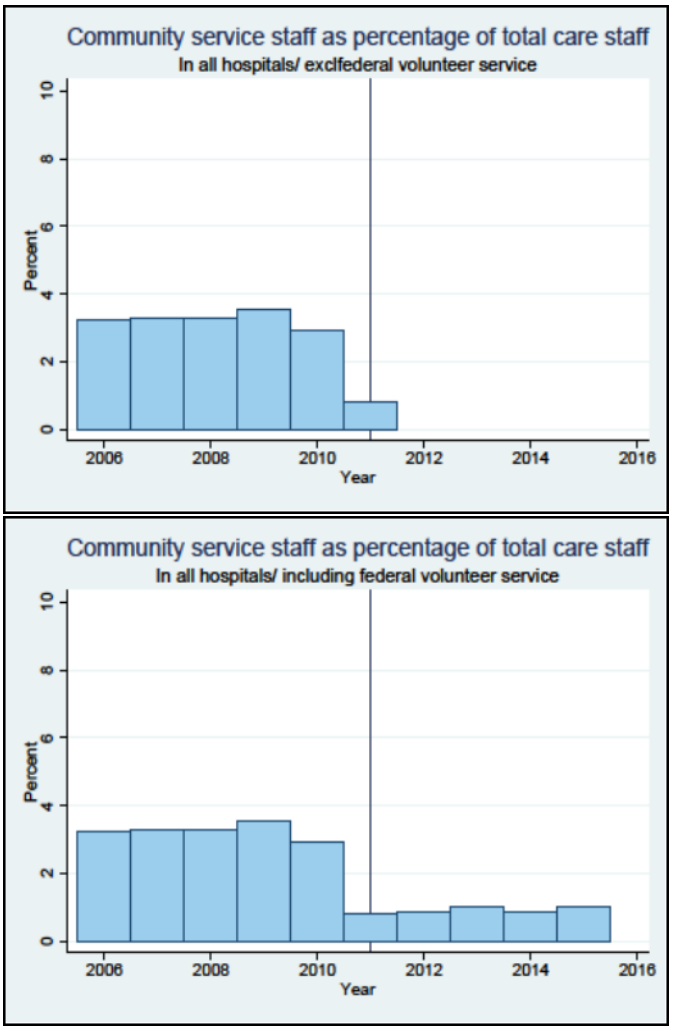
Control variables: As additional control variables I use hospital characteristics like type of sponsor (public, private, non-statutory welfare services), type of hospital (i.e. university hospital, hospital forming part of the German hospital plan, hospital with provision contract or others) and share of apprentices in total staff. Hospital size is controlled for by number of installed beds. Following the literature (compare, for instance, Breyer et al. 2013) these are summarized in 5 categories (less than 75, 75-124, 125-249, 250-499, 500

[44]Care hours equals reported care days x 24, whereas working hours of yearly average full-time equivalences equals FTE x 220 [working days per year] x 8 [hours].

or more). Additionally regional characteristics are accounted for by controlling for the degree of regional agglomeration by considering 8 categories ranging from *densely populated urban areas* to *rural county, lightly populated*. This is relevant because descriptive evidence has shown marked differences between treatment prevalence across counties. Hospitals with more than 100 beds also provide information on all of their patients, their main (and, if applicable, secondary) condition and additional details like age, gender, place of residency, day of hospitalization and discharge, ward with the longest duration of stay of the patient and if a surgery was performed. To additionally control for the type of patients hospitals treat, I include share of female patients, share of patients aged 75 or older and share of surgeries performed.

1.A.2 Additional Figures

Figure 1.7: Community Service in All German Hospitals



Source: German Hospital Statistics, Aggregate Data

CHAPTER 2

Free Universal Daycare: Effects on Children and Maternal Labor Supply

joint work with Christina Gathmann

2.1 Introduction

Many governments have invested sizable amounts of taxpayer's money into fostering daycare for preschool children.[1] Proponents of such policies argue that money invested in early childhood education is well spent as it simultaneously boosts female labor supply, benefits child development and promotes a level playing field for children from disadvantaged family backgrounds. Initially, policy-makers focused on providing daycare prior to school entry ("kindergarten"). Over time, political attention has shifted to expanding daycare for ever younger children.

Hitherto, the empirical evidence on daycare for older pre-school children is much more comprehensive than our understanding of the impact on children under the age of three.[2] How public daycare influences family choices and child development is likely dependent on the age of a child. Sensitive periods for certain skills like language acquisition and visual memory (see e.g. Robson 2002; Siegler et al. 2017) and possible dynamic comple-

[1] Countries like France, Sweden, Norway or Denmark have long offered universal access to public childcare. Others, like Germany, Spain, the UK or the U.S., have expanded public daycare and pre-K programs much more recently – beginning in the 1990s.

[2] A sizable literature investigates the labor supply effects of daycare prior to school entry (Gelbach 2002; Berlinski and Galiani 2007; Cascio 2009) and daycare for children between the ages of three and six (Schlosser 2006; Fitzpatrick 2010; Havnes and Mogstad 2011a; Nollenberger and Rodriguez-Planas 2015). A much smaller literature investigates the consequences of daycare attendance on the short- and medium-run development of 3-6-year old children (Datta Gupta and Simonsen 2010; Blanden et al. 2016; Cornelissen et al. 2018). Yet another literature analyzes more long-term consequences of daycare on educational attainment, for instance (see e.g. Black et al. 2014; Carneiro et al. 2015; Dustmann and Schönberg 2012; Havnes and Mogstad 2011b).

mentarities over time (Cunha and Heckman 2007; Heckman 2007) might make daycare more productive at some ages than others, for instance.[3]

Most previous evidence for very young children stems from highly targeted programs for disadvantaged families like Head Start, for instance.[4] Universal early childcare programs that offer public daycare to young children likely have different effects than targeted programs to help disadvantaged children (Baker 2011; Cascio 2017; Kottelenberg and Lehrer 2017). It is only very recently that a few studies on the consequences of public childcare for very young children on labor supply (Goux and Maurin 2010) and first skill development (Felfe and Lalive 2018; Fort et al. 2019) emerged.

In this paper, we exploit the staggered introduction of free universal daycare in German states to track how families with pre-school children between the age of 2 and 6 respond to and benefit from the policies. Specifically, we make four contributions to the literature on childcare and early childhood education. Most importantly, we can compare whether the margins of adjustment to and benefits of a free daycare policy differ for 2-year-olds and 5-year-olds, for instance. Since attendance rates in many countries have traditionally been much higher for older pre-school children, providing daycare free of charge is likely to generate stronger behavioral responses among families with very young children for which attendance rates are still low.

Second, we can assess the impact of free daycare on a range of family choices, including childcare arrangements and labor supply, as well as short-run child development. Analyzing the full range of family responses to daycare policies is crucial for interpreting the estimated effects on child development. Children are less likely to benefit from a free daycare policy, for instance, if parents switch from high-quality parental care to low-quality public care; they are more likely to benefit if families switch from low-quality informal (or parental) care to high-quality daycare instead.

Because the policies we study are universal, our third contribution is to shed light on who responds to the policy; and whether the benefits accrue to the average child or are concentrated among children from disadvantaged families, for instance. Whether public daycare is able to level the playing field between poor and more affluent families is of central interest to policy-makers concerned about equality of opportunity in early childhood education.

Finally, the policy we analyze offered childcare subsidies rather than merely expanding daycare supply. Providing additional daycare slots expands the choice set of parents,

[3] Also, earlier evidence showed negative effects of maternal labor supply among very young children that turn positive when the child is two years or older (James-Burdumy 2005). Also, the trade-off between parental preferences, returns to work and childcare constraints might be very different for a 2-year-old than for a 6-year-old.

[4] See e.g. Currie and Thomas (1995); Garces and Currie (2002); Love et al. (2005); Ludwig and Miller (2007); Carneiro and Ginja (2014); Walters (2015).

which generates substitution effects out of care at home or informal care into formal childcare. Free daycare, in turn, is equivalent to a price decline of public daycare relative to other childcare options and, since it is uncompensated, there will be both income and substitution effects on childcare and labor supply choices. Hence, children might still benefit from the policy even though parents do not adjust their labor supply behavior, for instance.[5]

Closest to our analysis are two studies that also analyze the effect of childcare subsidies on maternal labor supply and child development (Black et al. 2014; Baker et al. 2008).[6] Black et al. (2014) exploit income thresholds for subsidies targeted at 5-year-olds in Norway to identify the effect on long-run child outcomes in junior high school and female labor supply. Our study can provide evidence of childcare subsidies across the full age range of pre-school children. A second advantage of our study is that we explore the heterogeneity of effects for disadvantaged families, which are often the focus of the political debate on early childhood education. Baker et al. (2008) study the introduction of the Family Policy in Quebec that expanded daycare supply and further offered generous subsidies for children between the ages of 1 and 5. One advantage of our analysis is that we can exploit free daycare policies in nine states rather than just one state. Furthermore, we explore the heterogeneity of free daycare policies for different child ages and for disadvantaged families. Both types of heterogeneity are important to understand who responds to and benefits the most from free daycare policies; and to inform policy-makers on the targeting of public daycare policies.

Germany provides an interesting case to analyze childcare policies. Federal and state governments in Germany spend a lot of public resources, about 200 billions Euros per year, on various family policy measures (Bonin et al. 2013). Yet, many women, despite having surpassed men in their formal education, still drop out of the labor force or work only part-time once they have children. Many people argue that affordable childcare, especially for children under three, is crucial to boost female labor force participation and promote economic self-sufficiency, especially for economically disadvantaged families (e.g. Attanasio et al. 2008).

Our empirical analysis uses birthday cutoffs for school entry to define eligibility of a child for a free daycare slot in a state and year. Eligibility thus depends on the state of residence, the child's birth cohort and age. Hence, we compare the choices of families

[5] Furthermore, the behavioral response might be larger or smaller in response to declining daycare prices than in response to the availability of daycare. The labor supply effect of a free daycare policy might be smaller if maternal labor supply is mainly held back by rationing of slots rather than high prices.

[6] Focusing on labor supply and reforms in childcare subsidies a few studies find no effect on female labor supply in Sweden (Lundin et al. 2008) but sizable effects on employment and hours worked in Canada (Lefebvre and Merrigan 2008). Offering a generous child benefit for newborns results in a substantial postponement of maternal labor supply in Spain, however (Gonzalez 2013).

with eligible children in states with a free daycare policy in place to the choices of families with children in the same birth cohort and age who live in states without a free daycare policy. Both graphical evidence and estimates from placebo reforms confirm the absence of differential pre-trends between families in treatment and control states. We further bolster our findings by performing a large number of robustness checks to rule out that differences in child age, the supply of daycare, local economic conditions or other policy features may explain our results. Our alternative specifications and informal validity tests yield estimates that are very similar to the baseline estimates.

Existing studies on Germany all focus on the rapid expansion of public daycare for children between the ages of 3 and 6. Bauernschuster and Schlotter (2015) find that maternal labor supply responded strongly to the better availability of daycare. More recent contributions show that the average effect of daycare is small but benefits are largest for minority children and those children least likely to attend daycare (Cornelissen et al. 2018; Kühnle and Oberfichtner 2017). There is only one prior study that analyzes childcare attendance under the age of three (Felfe and Lalive 2018). They find developmental benefits for some, but not all children.[7] Our paper differs from prior studies on Germany along three dimensions: first, we study policies that affect children between the ages of two and six. We can therefore compare which and how families with children at different ages adjust their behavior and who benefits from the policy. Second, we analyze a free daycare policy rather than the availability of daycare per se. As discussed above, the behavioral responses and overall effects we expect from a free daycare policy are likely to differ from expanding daycare availability. Finally, our policies were introduced in many states, which differ widely in their economic and social structure. Hence, our estimates capture the average treatment effect of a universal free daycare policy rather than pick up the effects of specific conditions or policy features in one locality or state.

Our analysis yields five main findings. First, the free daycare policy only affects the youngest children aged between 2 and 3. Access to a free daycare slot raises daycare attendance in that age group by 8.4 percentage points or 17% relative to the pre-policy period. We further observe a corresponding decline in exclusive care at home. Interestingly, the use of informal daycare by relatives, friends or neighbors for children in this age group actually goes up by 8.3 percentage points or 22% suggesting that formal and informal daycare are complements in our context.

Second, daycare attendance for older children (aged 3 and above) does not respond to the free daycare policy. In particular, we find no effect for the most common policy that adopted a free year of daycare prior to school entry ('kindergarten'). The main reason

[7] Gathmann and Sass (2018) in turn analyze the effect of a home care subsidy, which is equivalent to a price hike for public daycare, on families with 2 years-old children in East Germany.

is that daycare attendance for children aged 3 and above has been high (94.4%) even before the policies were introduced. For most families with 3-6 year-old children, the free daycare policy is just windfall income. Our results indicate that some of the additional income is used to purchase informal childcare.

Third, we observe substantial positive employment effects among mothers with 2-3 year-old children after the policy is adopted. Labor force participation increases by 7.7 percentage points or 17% relative to the pre-reform period. Mothers with older children (aged between 4 and 6) in turn increase their labor supply at the intensive margin – working more full-time and increase working hours by almost two hours per week (or 11%). Overall, these results imply that the additional income saved from the free daycare slot is not used to reduce female labor supply.

Fourth, the free daycare policy has few persistent effects on child development. For the youngest children, we observe no overall effect, though a negative effect on skills in daily activities (i.e. whether the child can use a spoon on its own, for instance) and a positive effect on social skills. For children aged 5 to 6, we find no overall effect and also no effect on sub-categories measuring behavioral problems. Hence, the policy seems to do no persistent harm, but also creates few benefits in terms of cognitive or non-cognitive skills for the average preschool child.

Finally, we document substantial heterogeneity in the treatment effect: poor and low-skilled households respond more to the policy than the average family. Children from low-skilled and poor households are more likely to attend public daycare and less likely to be cared for exclusively at home. Poor children in particular benefit from free daycare, which boosts their cognitive and non-cognitive skills. In contrast, we find no effect on female labor supply suggesting that either the returns to work (more) or labor supply elasticities more broadly are small for low-income and low-skilled mothers.

Overall, the free daycare policy substantially reduced the gap between rich and poor as well as between skilled and low-skilled households. For both groups, the pre-reform attendance gaps of around 13 percentage points declines by up to two-thirds after introducing free daycare. Children from less educated households especially, benefit from the policy in terms of their cognitive and non-cognitive skills. Hence, a free daycare policy, despite being a universal policy, contributes to better equality of opportunities thus leveling the playing field between children from more and less advantaged backgrounds.

This chapter proceeds as follows. Section 2 provides relevant background information on public daycare and the adoption of the state-wide reforms. Section 3 discusses the policy variation and our estimation approach, while section 4 introduces the data sources. We present our main results in section 5 and discuss a battery of robustness checks in section 6. Finally, section 7 concludes.

2.2 Institutional Background

2.2.1 Public Daycare in Germany

We now provide relevant background information on the childcare market in Germany and the adoption of free daycare policies. Daycare outside the home is supplied by either the municipalities or private, non-profit providers, mostly churches and non-statutory welfare services. Municipalities supply around one-third of the childcare slots, while private, non-profit agencies provide around two-thirds. Private, for-profit childcare providers cover only a very small share of the market - around 2% for children under 3 and 0.3% for children from 3-6 years of age (Berger et al. 2008).

Federal regulations explicitly define three goals of public daycare: providing care and custody for preschool children; advancing their social and non-cognitive skills; and fostering the children's education and learning. In practice, many different educational approaches (like Montessori, Waldorf etc.) exist side-by-side. Most popular in center-based daycare is the situation-oriented approach, a social pedagogy tradition that stresses flexible schedules, problem-solving and social skills through play, social interaction and informal learning. This tradition contrasts with a more school-oriented approach that focuses on teaching cognitive skills and basic knowledge (Sohns 2009).

Germany's childcare system is considered of intermediate quality in terms of public expenditures, but of relatively homogeneous quality thanks to strict regulations of quality standards and high educational qualifications of childcare staff. Combined public and private expenditures on early childhood education are around 0.6% of Germany's Gross Domestic Product (GDP), which is similar to the EU average though below the expenditure share in France, the UK and some Scandinavian countries (OECD 2013). The federal and state governments put in place detailed regulations to ensure a certain quality level in daycare centers. All childcare facilities require a permit which may be revoked if standards regarding group sizes, educational background of the staff, the physical environment and standards for hygiene and security are not met. Even private, for-profit childcare providers comply with these regulations as they would otherwise not obtain the generous public subsidies that cover most of the facility's variable costs. The local and state youth offices are responsible to monitor the requirements and impose sanctions in case of non-compliance, up to the point of closing a facility.

The educational standard of childcare staff is high in international comparison. Each facility must have at least one professionally trained educator. Training as a child educator involves two years at a vocational school in combination with practical training followed by one year of practical training in a childcare facility. Many of the head teachers have

a diploma in social pedagogy or related subjects involving a curriculum of 3-4 years at a technical college with a focus on early childhood education. Aggregate data illustrate the high educational qualifications of childcare staff: 64% of all employees and 90% of those leading a group have obtained vocational training as an educator (OECD 2017). Regulations in each state further regulate group sizes with a maximum of 25 children. The actual child-staff ratio is with 12 children much lower (OECD 2013).

The period we study saw some changes in the availability of public daycare slots as shown in Figure 2.2 in the appendix. For 3-6 year-old children, the supply of public daycare hovers around 100% (see left y-axis) and does not change much over time.[8] The situation is different for children under the age of 3 where traditionally few slots were available. Starting in the early 2000s, the federal government has invested substantially in expanding the number of daycare slots for children under the age of 3.[9] As a result, childcare slots increased from under 10% in 2002 to more than 30% in 2015 (shown on the right y-axis in Figure 2.2). In the empirical analysis below, we will use district-level data on the supply of childcare slots to check that changes in the availability of daycare slots do not drive our results.

The expansion of supply might have potentially negative effects on daycare quality by increasing child-staff ratios or the number of staff without proper educational qualifications, for instance. Aggregate statistics show no change in the staff qualifications over time, however: the share of childminders with at least vocational training in early childhood education remains stable at 80% throughout. The quality might also suffer if group sizes per child minder increase in areas with large expansion of childcare slots. Yet, Figure 2.3 in the appendix shows little evidence for a worsening of child-staff ratios between 2006 and 2014. If anything, the number of children per child minder falls over time with an average of about 8 children aged between 3 and 6 and around 4 for children under the age of 3.[10] Overall, changes along the quality dimension do not seem to be a major concern for our analysis.

[8] Since 1996 federal regulation grants a daycare slot to 3-6 years-old children in all states.

[9] In 2008, the federal government decided to offer a daycare slot for all children after their first birthday from 2013 onwards. As a result, the supply of childcare, esp. for children under the age of 3, has grown substantially over time.

[10] Furthermore, regressing the group size (in $t + 2$) on the adoption of a free daycare policy for the specific age group (in t), state and year effects shows an increase in the group size in treatment states by around 0.5-0.7 children for 2-3-year-olds with no effect for older children. If parents value quality measured as small group sizes, these results suggest that our estimates are, if anything, a lower bound, especially for very young children.

2.2.2 Parental Fees and the Adoption of Free Public Daycare

Public daycare is heavily subsidized in Germany.[11] Parental fees cover less than 20% of the variable costs with the remainder being financed by state and local government funds (OECD 2017; Leu and Schilling 2008). Fees are typically set at the municipal level, which creates substantial variation in daycare prices both within a state and across states.[12] A typical range is between 0 and 220 euros per month for a part-time slot in daycare with fees increasing in parental income. Fees for a full-time slot in daycare can be as high as 800 euros per month for high-income parents in urban areas.[13] On average, parents in our data pay around 90 euros per month for a childcare slot between 2002 and 2014. Similarly, parents surveyed in the NEPS, a large panel study covering preschool and school children, also report paying around 86 euros per month for a childcare slot in 2011.

Between 2000 and 2016, nine states in West Germany introduced public daycare slots free of charge to eligible children.[14] Two West German states in turn have never offered free childcare over our sample period. Table 2.1 provides an overview of the implemented reforms: states differ both in the timing of policy adoption and how comprehensive the reforms are. Six of the nine states abolished parental fees only for the last year of daycare prior to school entry ("kindergarten") - when the child is 5 or 6 years old. Three states, which make up roughly 15% percent of our sample, introduced more comprehensive reforms. *Berlin*, for instance, offers free public daycare for all children aged between 2 and 6 since 2016. The policy was initially adopted in 2007 for the last year of daycare prior to school entry, then extended to two years of daycare in 2010, further expanded to three years of public daycare in 2011 and to four years in 2016. *Rhineland-Palatinate* phased in free daycare for all preschool children from 2-6 years of age between 2007 and 2010. *Hamburg*, in turn, abolished parental fees for the last daycare year in 2009 and extended the policy to all children aged 2 and above in 2014.

[11] Public daycare includes facilities for preschool children mostly provided by municipalities or private, non-profit providers like churches or welfare services. The share of private, for-profit providers is very low. Even private providers comply with state daycare regulations; otherwise, they would lose the very generous public subsidies which cover around 80% of the facility's variable costs.

[12] Unfortunately, there are no data sources that allow to trace daycare prices over time or their variation across space. As a general rule, the cost of a daycare slot to parents varies with the number of children in the household and parental income (Goerres and Tepe 2013).

[13] Expenditures for formal childcare are tax-deductible up to a limit of 4,000 euros per year; hence, net expenditures for childcare after taxes are somewhat lower.

[14] We focus in our analysis on the eleven states in West Germany as childcare provisions and female labor supply still differ between East and West Germany.

Table 2.1: Introduction of Free Childcare in West Germany

	Broad Age Group Covered	Year Adopted	State of Adoption
Last Year of Public Daycare	Ages 5-6	2000	Saarland
		2007	Rhineland-Palatinate
		2007	Berlin
		2007	Lower Saxony
		2008	Hesse
		2009	Hamburg
		2009-2010	Schleswig-Holstein
		2011	North-Rhine Westphalia
		2013	Bavaria
2nd Year of Public Daycare	Ages 4-5	2008	Rhineland-Palatinate
		2010	Berlin
		2014	Hamburg
1st Year of Public Daycare	Ages 3-4	2009	Rhineland-Palatinate
		2011	Berlin
		2014	Hamburg
Public Childcare (pre-K)	Ages 2-3	2010	Rhineland-Palatinate
		2016	Berlin
		2014	Hamburg
No Free Daycare Policy	Ages 2-6		Baden-Wuerttemberg
			Bremen

Notes: The table shows which states adopted free childcare in which year and for which broad age group of children. Schleswig-Holstein abolished free childcare in July of 2010. In Hamburg and Schleswig-Holstein, for instance, access to free childcare applies to a part-time childcare slot (up to 5 hours a day). In other states, free childcare applies to a slot up to 12 hours per day (full-time slot).

2.2.3 Determinants of Adoption

Our estimation strategy requires that the reforms and their timing have to be unrelated to female labor supply and childcare. We further want to rule out that omitted variables, like voter preferences, for instance, account for both the free daycare policy and family choices.[15] The political discussion prior to the introduction of free childcare in the nine states stressed equity concerns. The main concern was to provide access to early childhood education for all preschool children - independent of their family background and parental resources.[16] The political and media discussion does not indicate, for instance, that the reforms were implemented in order to increase female labor supply or to assist children lagging behind in their cognitive development.

Table 2.2 investigates the adoption decision more systematically: the dependent variables are indicators whether a state adopts any free childcare policy in year t (in columns (1)-(3)), and whether a state adopts a comprehensive free daycare policy in year t (in columns (4)-(6)). The explanatory variables are lagged two years and include basic socio-economic conditions (unemployment rate, GDP per capita, population, the shares of medium- and high-skilled employees and the share of women in the labor force), state and year fixed effects. The second specification (in columns (2) and (5)) adds the number of slots available per 100 children separately for children under 3 and children between 3 and 6. The third specification (in columns (3) and (6)) further controls for the vote share of conservative and left-wing parties in state elections to capture voter demand for free daycare policies.

Table 2.2 shows four interesting patterns: first, states with higher unemployment rates are less likely to adopt any or a comprehensive free daycare policy.[17] High unemployment rates reduce a state's financial capacity because of higher welfare payments and lower tax revenues. Below, we control for the unemployment rate and GDP per capita to rule out confounding changes in local economic conditions. Second, the female share in the workforce is unrelated to the adoption of free childcare. Hence, any changes in female labor supply we might observe are indeed a consequence of the reform rather than its motivation. Third, states with a better supply of daycare slots, especially for children under age three, are more likely to adopt a free childcare policy.[18] Responses to a free

[15] One might think that a free childcare policy is more likely on the agenda of a left-wing government. Yet, six states were governed by a conservative state government when they adopted a free childcare policy.

[16] See State Parliaments of *Berlin* (State Parliament Papers No. 16/2758 from November 10, 2009) or *North-Rhine-Westphalia* (State Parliament Papers No. 15/1929 from May 10, 2011) for two examples.

[17] An increase in the unemployment rate within a state by one standard deviation (or 1.36%) reduces the likelihood of adopting any free childcare policy by 23% (based on column (3)) and a comprehensive free childcare policy by 29% (based on column (6)).

[18] An increase in the supply of slots for children under 3 within a state by a standard deviation (5.38

daycare policy might be more pronounced if supply is readily available. We show in Section 2.6.1 below that estimates become larger when we control for the local supply of daycare slots.

Finally, there is no systematic relationship between electoral preferences and adopting a free daycare policy (see column (3)). This null effect reduces concerns that a shift in voter preferences in the years prior to the reform can account for both the policy and changes in family choices. Yet, a stronger left-wing vote share encourages the adoption of a comprehensive childcare reform (see column (6)).[19]

slots per 100 children) increases the likelihood of adopting any free childcare policy by 28% percent (based on column (3)).

[19]Raising the vote share for left-wing parties within states by one standard deviation (5.38 percentage points) increases the likelihood of adopting a comprehensive free childcare policy by 20% (based on column (6)). Supplementary regressions indicate that controlling for the vote shares in state elections as a proxy for electoral preferences does not affect our results on childcare arrangement or female labor supply (not reported), however.

Table 2.2: Determinants of Policy Adoptions

	Adopt Any Free Childcare Policy (9 out of 11 states)			Adopt Comprehensive Reform (3 out of 11 states)		
	(1)	(2)	(3)	(4)	(5)	(6)
Unemployment Rate (%)	-0.102	-0.148**	-0.169**	-0.181***	-0.165***	-0.217***
	[0.064]	[0.065]	[0.067]	[0.061]	[0.061]	[0.057]
GDP per Capita (Euros)	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000*
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
State Population (in 100,000)	-0.138***	-0.187***	-0.194***	-0.203***	-0.223***	-0.198***
	[0.043]	[0.045]	[0.046]	[0.045]	[0.046]	[0.043]
Share Medium-Skilled Employees	0.138	-0.048	-0.086	0.256***	0.152	0.123
	0.088]	[0.098]	[0.105]	[0.0874]	[0.097]	[0.097]
Share High-Skilled Employees	0.179**	0.043	0.021	0.323***	0.165*	0.116
	[0.080]	[0.088]	[0.090]	[0.080]	[0.089]	[0.085]
Women in Workforce (%)	0.079	0.055	0.044	0.074	0.113	0.029
	[0.089]	[0.088]	[0.092]	[0.085]	[0.085]	[0.083]
Slots for Children Aged 3-6 (per 100 children)		0.014**	0.015**		0.022***	0.018***
		[0.006]	[0.007]		[0.007]	[0.007]
Slots for Children Under 3 (per 100 children)		0.045***	0.052***		0.021	0.018
		[0.016]	[0.018]		[0.017]	[0.017]
Conservative Vote Share in State Elections (%)			0.001			0.012*
			[0.008]			[0.007]
Left-Wing Vote Share in State Elections (%)			0.002			0.027***
			[0.010]			[0.009]
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	130	130	122	120	120	112
R Squared	0.691	0.726	0.739	0.711	0.746	0.794

Notes: The dependent variable in columns (1)-(3) is an indicator equal to one if a state has adopted any free childcare policy in year t and zero otherwise; in columns (4)-(6), the dependent variable is equal to one if a state has adopted a comprehensive reform where preschool children are eligible for multiple years of free daycare. The sample consists of all West German states including Berlin over the period 2000-2014. All independent variables are lagged two years. Vote shares are taken from state election results and assigned the value of the last state election in non-election years. In addition to the variables shown in the table, the specifications also include state and year fixed effects. *** p<0.01, ** p<0.05 and * p<0.1.

Sources: Aggregate Statistics from the Federal Statistical Office, Social Security Data and German Youth Office

2.3 Empirical Strategy

2.3.1 Sources of Variation Induced by the Reforms

The free daycare reforms create three sources of variation for our empirical analysis: which states implemented a reform; which birth cohorts are eligible for free daycare; and the age groups covered by a reform. The first source of variation is straightforward: nine states adopted a reform, while two did not adopt any. As the nine states adopted the policy in different years ranging from 2000 to 2016, a child's eligibility further depends on its birth cohort.[20] Finally, the reforms cover children in different age groups. Nine states adopted free daycare for the last year before school entry ("kindergarten"), while three states adopted more comprehensive reforms covering younger children ("pre-K") as well. States define eligibility for free daycare based on birthday cutoff rules, which are also used to determine school entry. The last year of public daycare, for example, is defined as the 12 months preceding the school year in which the child turns six before the cutoff month. School, and thus the daycare year, typically start in August and last until July of the following year.[21] Hence, a child born in June of 2003, for instance, enters school in August of 2009 and thus starts its last year of daycare ("kindergarten") in August of 2008.[22] That same child would enter the first year of public daycare, which typically starts at age 3, in August of 2006. Eligibility for the youngest group of children (aged between 2 and 3) is determined by their second birthday. Accordingly, we use the birth date range from a child's second birthday to the time it is predicted to entry into public daycare to define eligibility for a 2-year-old.[23]

[20] The policies are adopted in January except in two states, which introduced it at the beginning of the school year (in August or September). The timing of adoption will not affect our estimates as children in the last year of childcare, for instance, are in these cases eligible for the period from January to August only (rather than the full year from August to July). As long as parents know about the adoption of the policy by the time children typically enroll in daycare, the timing of adoption does not affect our estimation strategy.

[21] There is some variation as most states engage in rotating summer breaks of six weeks starting as early as late June and as late as early August. As each state will start the summer break early in some years and later in others, this rotating scheme will, if anything, introduce classical measurement error in our estimation.

[22] It is important to stress that parents can actually enroll their child in daycare at any time during the year. In practice most children start public daycare at the beginning of the school year when daycare slots become widely available. Kühnle and Oberfichtner (2017) show that over 70% of children enter daycare when the school year starts in August or September. The remaining children enter at the 10 other months of the year roughly in the same proportion. Earlier or later entry into daycare will not invalidate our estimates as it only implies that we do not observe a child in daycare at the beginning of the daycare year (in the case of late entry); or that we observe a child attending an even earlier daycare year (in the case of early entry).

[23] In *Hamburg* children are eligible once they turn one. Yet, we have only about 10 children in our data which would be affected by this policy. We thus focus in our analysis on preschool children aged 2 and above.

States differ in the cutoff month they apply: five states use June 30, five states September 30 and one state uses December 31 as the cutoff rule. We therefore define a child's re-centered birth cohort as the 12 months following the cutoff. For a child living in a state with a June 30 cutoff, we define the re-centered birth cohort of 2003 from July 1, 2002 to June 30, 2003. Similarly, the 2003 birth cohort in a state with a September 30 cutoff is defined as children born between October 1, 2002 and September 30, 2003.

Defining birth cohorts in this way ensures that all children of a certain birth cohort are supposed to enter school and daycare in the same year: the 2003 birth cohort enters school in August of 2009, the last daycare year in August of 2008, the first (of three) daycare year in August of 2006 and so on irrespective of the state's cutoff month. In treatment states, the birth cohort defines the set of eligible children. *Hamburg* introduced a free last daycare year in 2009, for instance. Hence, the first birth cohort eligible for this policy is the 2004 birth cohort (defined as those born between July 1, 2003 and June 30, 2004). In control states, the birth cohorts identify the set of children who would be eligible if the state introduced a free daycare policy for the same age group.

A potential disadvantage of defining birth cohorts in this way is that children belonging to the same birth cohort differ in their actual age (by up to 6 months) at a given point in time.[24] Such age differences might affect daycare and labor supply choices as well as a child's cognitive skills even independently of the reforms (see e.g. Black et al. 2011: for evidence from Norway). In the robustness section below, we show that controlling for the age of the child in 3-month intervals does not affect our results.

To capture the third source of variation, we use the broad age group of a child (ages 2-3, 3-4, 4-5 and 5-6), which characterizes the daycare year a child would typically attend. We use this variation in two ways: to control in the estimation for state-level differences in daycare attendance by age groups and for any differences in attendance across birth cohorts. More importantly, we use the broad age group of children to explore whether a free daycare policy has different effects for older children (5-6 years of age) compared to younger children (2 years of age, for instance). We now discuss our estimation strategy.

2.3.2 Estimation Strategy

We start with an analysis of access to a free daycare year prior to school entry ("kindergarten"), the most common policy adopted. Hence, we restrict the sample to children

[24] Children in states with a cutoff rule in June, for instance, will be slightly older when they enter their last daycare year than children in states with a September or December cutoff rule. A regression of the cutoff month on whether a state adopts any free childcare and year dummies shows that treatment states have their cutoff date somewhat earlier in the year; hence, children in treatment states are slightly older when they enter their last daycare year than children in control states.

whose age group typically attends the last year of daycare in their state. In particular, we estimate variants of the following model:

$$Y_{iacs} = \beta * Eligible_{cs} + \gamma' X_{iacs} + \alpha_s + \theta_c + \varepsilon_{iacs} \tag{2.1}$$

where Y_{iacs} represents outcome of child (or parent) i of birth cohort c in state s. Our main outcomes are childcare choices, maternal labor supply as well as child cognitive and non-cognitive skills. The key independent variable $Eligible_{cs}$ is equal to one if a child is eligible for a free daycare slot and zero otherwise. As discussed in the previous section, eligibility depends on whether the state of residence has adopted a free last year of daycare and whether the child's birth cohort is eligible for a free slot.

All specifications further include state (α_s) as well as birth cohort (θ_c) fixed effects to allow for differential childcare attendance across states and differential trends in daycare attendance for earlier and later cohorts. We further include additional variables X_{iacs} to control for family-level differences and improve the precision of estimates: child gender, the parent's education, age and marital status, whether the parent is foreign-born, household size, the number of dependent children and number of infants. To adjust for changes in local economic conditions, we control for state GDP per capita and unemployment rate. Our results do not depend on the specific set of controls (as shown in Section 6.1. below). The key parameter of interest, β, then identifies the Intention to Treat Effect (ITT) effect of being eligible for free kindergarten relative to children of the same birth cohort in the control states, which combines never adopting and later adopting states.

We then investigate how free daycare affects family choices for all preschool children aged between 2 and 6. Here, we estimate variants of the following model:

$$Y_{iacs} = \beta * Eligible_{acs} + \gamma' X_{iacs} + \alpha_s * \lambda_a + \theta_c * \lambda_a + \varepsilon_{iacs} \tag{2.2}$$

where Y_{iacs} are the same outcome variables as in equation 2.1 above. The treatment variable $Eligible_{acs}$ is now equal to one if a child in a certain age range and birth cohort is eligible for a free daycare slot in a treatment state. We include the same additional controls (X_{iacs}) as above.

We further include a full set of state x broad age group ($\alpha_s * \lambda_a$) fixed effects to absorb any differential trends in age-specific childcare and maternal labor supply choices across states. If families in *Hamburg* send their child to daycare earlier than in *Bavaria*, for instance, even independently of any free daycare policy, these differential choices will be fully absorbed by the state x broad age fixed effects. Finally, we include a full set of birth cohort x broad age group fixed effects ($\theta_c * \lambda_a$) to capture any changes in childcare choices between earlier and later birth cohorts. These fixed effects account for differential age of

entry patterns, e.g. if the age of entry into daycare declines for later birth cohorts, for example. In variants of equation (2.2), we also allow the effect of the treatment variable to vary with the broad age range of the child by replacing $Eligible_{acs}$ with $Eligible_{acs} * \lambda_a$. The parameters of interest β in (2.2) are then identified by comparing changes across birth cohorts for eligible children residing in a treatment state to the changes for children in the same broad age group in a control state. Given the fixed effects included in the model, the parameter is identified from the triple interaction between state, birth cohort and broad age group.[25]

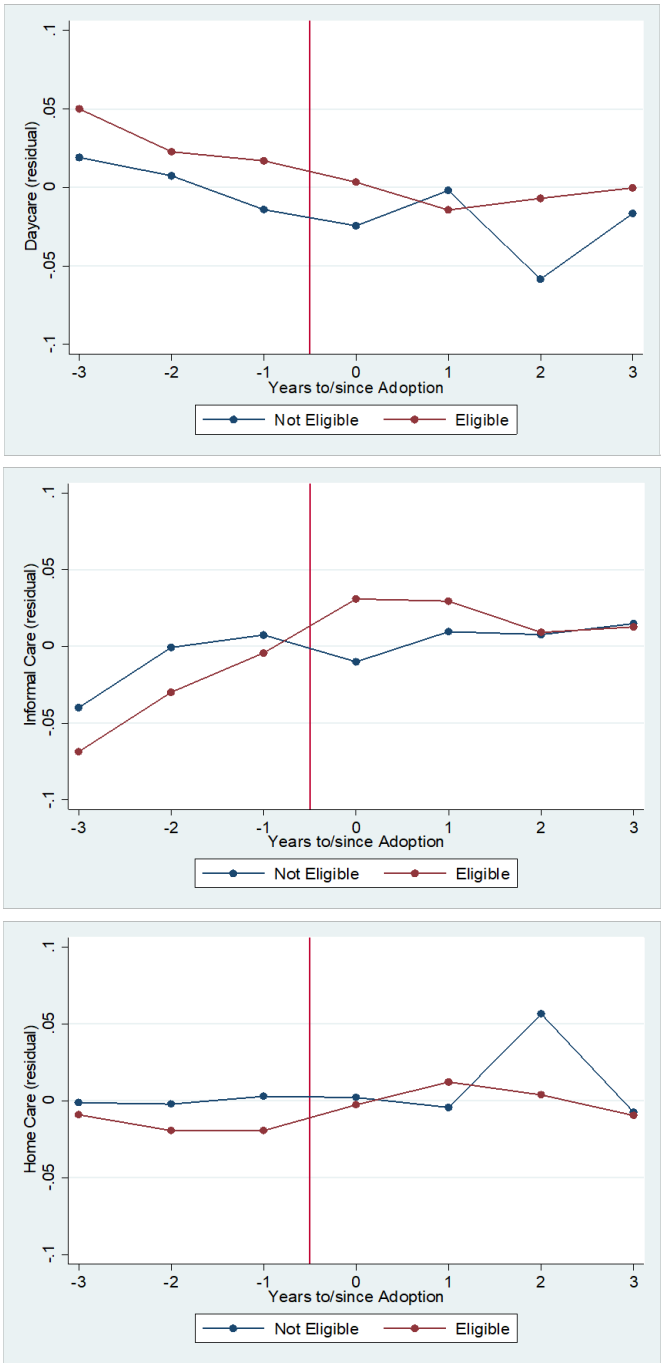
As with any difference-in-differences strategy, causal interpretation relies on the assumption that in the absence of reforms, the evolution of outcomes would be parallel for the same birth cohorts across states regardless of whether and when they implemented a reform. While we cannot directly test for this assumption, we provide graphical evidence that outcomes evolve similarly in the pre-reform period. Using residual choices of childcare obtained by estimating equation (2.2) without the $Eligible_{acs}$ variable (but including all other control variables), Figure 2.1 plots residual means for the three years preceding and following free daycare reforms (where the implementation of the first reform in a state is defined as year 0). Control states combine never adopters and later adopting states. For never adopters, we assign the mean reform year among treatment states as year zero. The graphs show that family choices move in parallel in treatment and control states prior to the reforms.

We further bolster the validity of our identification strategy using placebo reforms and a range of alternative specifications in Section 2.6.1. We find very similar estimates if we only use a narrow range of birth months around the cutoff date for school entry – an estimation strategy commonly used in the school entry literature.[26] Another concern with difference-in-differences analysis is the correct computation of standard errors. To account for within state dependence, our baseline estimations cluster standard errors at the state level (Bertrand et al. 2004; Cameron and Miller 2015). In Section 2.6.3 below, we demonstrate that using alternative estimators for the variance-covariance matrix (no clustering, less aggregate clustering or accounting for the small number of clusters) does not affect our inference.

[25] A regression of the eligibility variable on all control variables and fixed effects yields a R^2 of 0.69. As such, there is a lot of variation left in the treatment variable to identify the effects of the free daycare policy.

[26] In principle, one could also use the cutoff rules and birthday information to implement a RDD design for estimation (as in Fitzpatrick 2010; Gormley and Gayer 2005). In practice, our sample sizes are too small for such a data-intensive procedure, however.

Figure 2.1: Care for Eligible and Non-eligible Children



Top: Daycare Attendance, Center: Informal Care, Bottom: Exclusive Home Care

2.4 Data Sources

2.4.1 The Socio-Economic Panel

Our empirical analysis uses data from the Socio-Economic Panel (SOEP 2017), which surveys around 9,000 West German households each year about their childcare choices, labor supply and income. We restrict our sample to the period from 2000 to 2016, which includes at least six pre-policy years and up to ten years after a reform.[27] We include in our sample all families in West Germany with at least one preschool child aged between 2 and 6. The data appendix provides more details about our sample and the variables used in the empirical analysis.

Parents report whether their children attend public daycare, whether people from outside the household (e.g. relatives, friends, neighbors or a child minder) care for the child or whether childcare is exclusively provided by members of the household instead. Note that home care does not imply that all care is provided by the parents; it also includes childcare by other household members like older siblings, au pairs or grandparents. Based on this information, we code an indicator variable whether a household uses public daycare or not, whether the household uses informal care (as alternative or in addition to public daycare) and whether the household does not use any childcare outside the home. To test whether access to free childcare encourages mothers with preschool children to enter the labor market or work more hours, we use information on labor force participation and working hours of the responsible parent.[28] Employment here comprises full- or part-time employment, employment for less than 400 euros per month (which is exempt from social security contributions) and vocational training. Mothers on parental leave are considered not employed. Working hours are contractual hours and measured per week. We use a number of socio-demographic characteristics of the child, the responsible parent and the household to control for other influences on childcare arrangements or labor supply. Finally, we merge our data on families with preschool children from the SOEP with administrative data on the supply of public daycare slots from the Child and Youth Services at the district level and with data on state-level unemployment and GDP per capita from the Federal Statistical Office.

Table 2.10 shows descriptive statistics for our sample of families with preschool children in West Germany separately for the pre-reform period (2000-2006) and the post-reform period (2007-2016). Around 80% in our sample of preschool children between 2 to 6 attend public daycare but most children attend for less than 8 hours per day. Informal

[27] While the *Saarland* introduced a free daycare year in 2000, this state has less than one million inhabitants and constitutes less than 1% of our sample of preschool children.

[28] The responsible parent is the mother (99%) or another female adult like the grandmother.

care is still common and often combined with public daycare in order to cover childcare needs. Maternal employment is with roughly 50% relatively low compared to the United States, for example; and most working mothers work part-time, i.e. less than 30 hours per week. Figure 2.1 shows that daycare attendance has been rising mostly for children between the ages of 2 and 4. Attendance rates are high for older children (aged 5-6) even at the beginning of our sample period and do not change much over time.

2.4.2 Supplementary Information on Child Outcomes

To analyze child outcomes for 2-3 year-old children, we use a supplementary questionnaire of mothers.[29] Four skill categories are surveyed using an adapted Vineland Adaptive Behavior Scale (VABS): motor skills, language ability, social skills and skills in daily activities (see Sparrow et al. (2005) for more details). Five questions are used for assessment in each category, e.g. whether the child can form a sentence with multiple words (for language skills) or draw recognizable figures (for motor skills). For each question, the mother reports whether the child is able (2 points), not able (0 points) or only partially able (1 point) to perform a particular task. We construct a score for each category (language, motor skills etc.) by summing the responses to the individual items. We further calculate a total VABS score across all four categories ranging from 0 to 40 (mean: 28.5, standard deviation: 8.2 in our sample). Finally, we standardize the score to have mean zero and standard deviation of one. A larger score implies that a child is better able to perform the specific tasks.

To assess child outcomes for older children, we use a questionnaire of mothers with 5-6 year-old children. Here, child outcomes are measured by an adapted version of the Strengths and Difficulties Questionnaire (SDQ) by Goodman (1997). Mothers assess emotional and conduct problems, hyperactivity/inattention and peer relationship issues of their child relative to other children in the same age range. The four dimensions are summed to a total SDQ score ranging from 0 to 23 (mean: 6.1, standard deviation: 4.1 in our sample). As for younger children, we standardize the total score and the sub-scores to have mean zero and standard deviation of one. Larger values indicate *more* behavioral problems. We now turn to our main results.

[29] Parental assessments, often the only source of information on very young children, may suffer from systematic biases. In appendix 2.A.3, we assess the relationship between parental assessments and childcare choices in more detail. The results in Table 2.11 indicate that maternal assessments do reflect actual changes in skills rather than biased parental perceptions.

2.5 Empirical Results

2.5.1 Childcare Arrangements

We first study how free daycare affects childcare choices in the last year before school entry ("kindergarten"). Results from estimating linear probability models of equation (2.1) on children in their last childcare year (as defined by the school entry rules) are reported in Table 2.3. The dependent variables are binary indicators whether a child attends public daycare (columns (1)-(2)), informal childcare by friends, relatives, neighbors or a child minder (columns (3)-(4)), or whether the child is exclusively cared for at home (columns (5)-(6)) respectively. The last two columns (columns (7)-(8)) report whether a child attends daycare full-time (conditional on being in public daycare). The main independent variable is equal to one if a child has access to a daycare slot free of charge and zero otherwise.

Table 2.3 reveals no behavioral responses to the free daycare slot for children between the ages of 5 and 6: there is neither an increase in attendance for the free public daycare nor any substitution patterns from or to other childcare modes.[30] Why do we see no behavioral response to the free daycare year? The last row shows that almost all (97%) children between the ages of 5 and 6 attend daycare even prior to the reforms. Hence, there was very little room for raising daycare attendance. As such, offering the last daycare year free of charge is mostly windfall income for the average family with eligible children.

Next, we turn to the reform effects for younger, pre-K children. The sample now includes all preschool children aged between 2 and 6. The dependent variables in Table 2.4 are again binary indicators equal to one if the family uses a certain childcare mode and zero otherwise. We now estimate linear probability models according to equation (2) where the main independent variable is equal to one if a child in birth cohort c living in state s is eligible for free daycare in the broad age range a. In addition to the control variables in Table 2.3, we add a full set of state x broad age range and birth cohort x broad age range fixed effects. The second specification (in even columns) interacts the age group with eligibility to allow reforms effects to vary by child age.

We find few effects on childcare choices on average: there is little effect on public daycare or exclusive home care (see columns (1) and (5)) but a positive, though not significant effect on informal daycare (see column (3)). Allowing the treatment effects to vary across child age, we see that the reforms encourage earlier entry into daycare: attendance for children

[30] The number of observations is lower for informal and home care because we have no information whether a household uses informal childcare in 2003. Furthermore, information on full-time attendance is available only until 2009.

between the ages of 2 and 3 increases by 8.4 percentage points (see column (2)). Compared to the mean attendance of 51 percentage points in the pre-policy period, the reform effect is with 17% quite large. In contrast, there is no change in childcare arrangements of children aged between 3 and 6 – likely because most children (94%) attend childcare in the pre-reform period. For the youngest children, public daycare attendance also grows at the intensive margin, though the coefficient on full-time attendance is not statistically significant (see columns (7) and (8)).

The decline in exclusive care at home is the mirror image of the changes in daycare attendance (see column (6)): 2-3 year-old children are 7.9 percentage points less likely to be exclusively cared for at home. Relative to the 30 percentage points cared for at home in the pre-policy period, the effect amounts to a decline of 27%. As for public daycare, we find no responses for older preschool children. Interestingly, informal daycare increases for all children between the ages 2 and 5 by 8.3 percentage points (see column (4)), but not for children just prior to school entry (between the ages 5 and 6). These results suggest that public daycare and informal childcare are complements for the youngest children, while income effects account for the increase among older preschool children.[31]

2.5.2 Maternal Labor Supply

Access to free childcare could boost the labor supply for mothers of pre-school children if returns to work increase, e.g. by reducing fixed costs of work; it could reduce maternal labor supply if the additional income is used to buy maternal leisure. Table 2.5 investigates labor force participation (in columns (1)-(2)), whether the mother works full-time (in columns (3)-(4)) and the number of contractual working hours (in columns (5)-(6)) based on estimating equation (2.2) and the same control variables as in Table 2.4. The second specification (in even columns) interacts the broad age range with eligibility to allow reforms effects to vary by the child's age.

We find few responses in maternal labor supply on average along the extensive (column (1)) and the intensive margin (columns (3) and (5)). Together with the evidence on childcare choices (see Table 2.4), these findings indicate that parents use the additional income (not spent on daycare fees) for informal childcare and activities other than parental working time.

Despite the zero effect on average, mothers with 2-3 year-old children increase their labor force participation by 7.7 percentage points or 17% (0.077/0.44). In contrast, there is little change in maternal labor supply along the intensive margin for this group. Moth-

[31] Evidence for East Germany (Gathmann and Sass 2018) suggests that public daycare and informal care are complements for very young children aged between 2 and 3, as daycare is often part-time or the hours are not flexible enough to cover a full workday.

Table 2.3: Free Last Year of Public Daycare and Childcare Arrangement

	Public Daycare Ages 5-6		Informal Childcare Ages 5-6		Exclusive Care at Home Ages 5-6		Daycare Fulltime Ages 5-6	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Eligible for Free Daycare	-0.007	-0.014	-0.019	-0.017	0.008	0.013	0.069	-0.004
	[0.008]	[0.014]	[0.014]	[0.018]	[0.007]	[0.011]	[0.068]	[0.075]
Parental Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Household Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Birth Cohort Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State-specific Cohort Trends	No	Yes	No	Yes	No	Yes	No	Yes
Observations	4,402	4,402	4,230	4,230	4,199	4,199	1,725	1,725
R Squared	0.028	0.032	0.067	0.068	0.030	0.037	0.149	0.167
Mean Outcome Pre-Policy Period	0.969	0.969	0.357	0.357	0.017	0.017	0.227	0.227

Notes: The table reports how childcare arrangements change if the last year of public daycare ("kindergarten") is offered free of charge. The sample is restricted to children in the last daycare year (i.e. 12 months prior to school entry). The dependent variables are all binary indicators. In columns (1) and (2), it is equal to one if a child attends public daycare and zero otherwise; in columns (3) and (4), it is equal to one if the family uses informal childcare by relatives, neighbors or friends for the eligible child and zero otherwise. The dependent variable in columns (5) and (6) is equal to one if the child is exclusively cared for at home and zero otherwise. In columns (7) and (8), the dependent variable is equal to one if a child attends daycare full-time and zero if it only attends in the morning or afternoon (conditional on attending daycare). The key independent variable "Eligible for Free Daycare" is equal to one if a child is eligible for free public childcare in the last year before school entry and zero otherwise. Eligibility depends on a child's state of residence, birth date and the cutoff rules for school entry (see main text for details). All specifications include household characteristics (household size, number of children and whether there is a child under age 1), parental characteristics (age, education and marital status of the responsible parent) and child gender, fixed effects for the state of residence and the birth cohort of the child as well as GDP per capita (linear and squared term) and unemployment rate (linear and squared term) in the state. Even columns also control for state-specific cohort trends. The last row reports the mean of the respective dependent variable in the pre-policy period (2000-2006). Standard errors are clustered at the state level. *** $p<0.01$, ** $p<0.05$ and * $p<0.1$.

Source: Socio-Economic Panel (2000-2016).

Table 2.4: Access to Free Public Daycare and Childcare Arrangements for All Children

	Public Daycare Ages 2-6		Informal Childcare Ages 2-6		Exclusive Care at Home Ages 2-6		Daycare Fulltime Ages 2-6	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Eligible for Free Daycare	0.013	0.084**	0.038	0.083***	-0.016	-0.079**	0.115	0.115
	[0.016]	[0.034]	[0.025]	[0.012]	[0.013]	[0.031]	[0.070]	[0.070]
Eligible*Ages 3-4		-0.086**		0.002		0.099***		-0.071
		[0.032]		[0.028]		[0.026]		[0.277]
Eligible*Ages 4-5		-0.061*		-0.003		0.065**		0.015
		[0.030]		[0.015]		[0.025]		[0.169]
Eligible*Ages 5-6		-0.053		-0.104***		0.077*		-0.110
		[0.051]		[0.025]		[0.039]		[0.177]
Parental Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Household Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Birth Cohort Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age Group Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State * Age Group FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Birth Cohort * Age Group FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	25,911	25,911	24,612	24,612	24,635	24,635	6,182	6,182
R Squared	0.360	0.360	0.066	0.067	0.211	0.211	0.114	0.114
Mean Outcome Pre-Policy Period	0.711	0.711	0.408	0.408	0.160	0.160	0.207	0.207

Notes: The table reports how the adoption of free public childcare for younger children affects childcare arrangements. The dependent variables are binary indicators: the dependent variable in columns (1) and (2) is equal to one if the child attends public daycare and zero otherwise, while in columns (3) and (4), it is equal to one if the parent uses informal childcare by relatives, neighbors and friends and zero otherwise. In columns (5) and (6), the dependent variable is equal to one if the child is exclusively cared for at home and zero otherwise; and in columns (7) and (8) it is equal to one if the child attends daycare full-time and zero if it attends part-time (conditional on attending daycare). The key independent variable "Eligible for Free Daycare" is equal to one if a child is eligible for free public childcare in its state of residence. Eligibility depends on a child's state of residence, birth date and cutoff rules for school entry (see main text for details). All specifications include household characteristics (number of children and whether there is a child under age 1), parental characteristics (age, education and marital status of the responsible parent), child gender, local economic conditions (GDP per capita and unemployment rate), state and cohort fixed effects as well as state x age group (2-3, 3-4, 4-5 and 5-6) fixed effects and cohort x age group fixed effects. The specifications in even columns (2), (4) and (6) interact the indicator for eligibility with the child's age group. The last row reports the mean of the respective dependent variable in the pre-policy period (2000-2006). Standard errors are clustered at the state level. *** p<0.01, ** p<0.05 and * p<0.1.

Source: Socio-Economic Panel (2000-2016).

ers of older preschool children (aged between 3 and 6) do not change their labor force participation but increase their hours of work. Mothers of 4-5 year-old children increase full-time work, while mothers of pre-school children (between the ages 5 and 6) increase their working hours both contractual and also actual hours worked. Contractual hours increase by 2.3 hours per week or about 11% (2.34/21.24).

Mirroring the effect on childcare choices, mothers of 2-3-year-olds indeed use a free daycare slot together with informal daycare to increase their labor supply. For older children (between the ages 3 and 5), some of the additional funds are used for informal care with no clear cut effects on maternal labor supply. For children just prior to school entry ("pre-K"), we find few changes in childcare choices, but an increase in working hours. Hence, for most of the children in our sample, the additional disposable income from abolishing daycare fees are not used for purchasing maternal leisure.

2.5.3 Short-Run Child Outcomes

We can further assess whether free daycare policies have any short-run consequences on child development up to age 6. For 2-3-year-olds, the dependent variables are now the Vineland Adaptive Behavior Scale (total score) as well as its sub-scores for motor skills, language skills, social skills and skills in daily activities. For 5-6-year-olds, the dependent variables are the SDQ (total score) as well as its sub-scores for conduct problems, emotional or peer problems and attention problems. The sample is much smaller than our main sample because skill development has only been assessed for children aged between 2 and 3 since 2005 and for children aged between 5 and 6 since 2008.

As before, we estimate models based on equation (2.2) where the eligibility variable is one if a child in a certain birth cohort and age range is eligible for free daycare in its state of residency; and zero otherwise. In addition to our main control variables (see Table 2.4 and Table 2.5), we also control for child age fixed effects (3 months window) and the survey month to adjust for age differences in cognitive and non-cognitive development. Each entry in Table 2.6 is an estimate of the treatment variable from a separate regression.

For children between the ages of 2 and 3, the left-hand side of Table 2.6 shows no effect on average child development. Looking at the individual subcategories, we observe a decline in skills in daily activities (like eating with a spoon correctly) of about 0.15 of a standard deviation. At the same time social skills seem to improve by 0.08 of a standard deviation (though the latter effect is not statistically significant). These results are in line with the fact that free daycare encourages earlier attendance in public daycare (see Table 2.4, column (2)): childminders are likely to have less time to teach each individual child how to use a spoon; at the same time, the child also spends more time interacting with other

Table 2.5: The Effect of Free Childcare on Female Labor Supply

	Labor Force Participation Ages 2-6		Work Full-time Ages 2-6		Contractual Working Hours Ages 2-6	
	(1)	(2)	(3)	(4)	(5)	(6)
Mother of Eligible Child	0.012	0.077***	0.029	0.020	1.037	-0.176
	[0.018]	[0.016]	[0.018]	[0.020]	[0.678]	[0.696]
Eligible*Ages 3-4		-0.057		-0.068*		-0.764
		[0.038]		[0.032]		[0.921]
Eligible*Ages 4-5		-0.100**		0.067**		1.565
		[0.035]		[0.025]		[1.299]
Eligible*Ages 5-6		-0.092*		0.018		2.343**
		[0.046]		[0.016]		[0.765]
Birth Cohort Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Age Group Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
State * Age Group FE	Yes	Yes	Yes	Yes	Yes	Yes
Birth Cohort * Age Group FE	Yes	Yes	Yes	Yes	Yes	Yes
Parental Controls	Yes	Yes	Yes	Yes	Yes	Yes
Household Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	26,329	26,329	13,867	13,867	11,481	11,481
R Squared	0.159	0.160	0.059	0.060	0.134	0.134
Mean Outcome Pre-Policy Period	0.460	0.460	0.178	0.178	20.51	20.51

Notes: The table reports how the adoption of free public childcare affects maternal labor supply. The dependent variable in columns (1)-(2) is a binary indicator equal to one if the mother is in the labor force and zero otherwise; in columns (3)-(4), the dependent variable is a binary indicator equal to one if the mother works full-time (i.e. more than 30 hours per week). The dependent variable in columns (5)-(6) is the number of working hours (in the contract) conditional on being employed. The key independent variable "Mother of Treated Child" is equal to one if the child is eligible for free public childcare and zero otherwise. Eligibility depends on a child's state of residence, birth date and the cutoff rule for school entry (see main text for details). Even columns interact the indicator for eligibility with the broad age range of the child. All specifications include state and birth cohort fixed effects as well as state x age range and birth cohort x age range fixed effects. All specifications control for economic conditions (state unemployment and GDP per capita), parental characteristics (age, education and marital status of the responsible parent), household controls (household size, number of children and whether there is a child under age 1) and child gender. The last row reports the mean of the respective dependent variable in the pre-policy period (2000-2006). Standard errors are clustered at the state level. *** p<0.01, ** p<0.05 and * p<0.1.

Source: Socio-Economic Panel (2000-2016).

small children in a daycare facility than with its parents, which could boost its social skills.

Are these effects temporary or do we still observe them for children between the ages of 5 and 6? The right-hand side of Table 2.6 suggests no statistically significant increase in behavioral problems among eligible children. The coefficient on the total score is very close to zero and not significant. Two sub-scores (for conduct and emotions) are negative suggesting fewer behavioral problems, while two others (for peers and attention) suggest more behavioral problems. None of the coefficients reach statistical significance, however. Based on the available data, a free daycare policy seems to have few longer-lasting negative consequences for children with some adverse immediate effects for very young children.

2.5.4 Heterogeneity Across Families

Policy-makers often favor subsidies for public daycare out of equity concerns (see Section 2.2.3). Access to free daycare with its trained educators, toys and a stimulating environment might boost skill development, especially for children from disadvantaged family backgrounds. Yet, do we actually see any leveling of the playing field for vulnerable subgroups like single mothers, poor or low-skilled households after the reforms?

Disadvantaged families might respond more to free daycare than the average family because parental fees constitute a larger share of their total household income, for instance. At the same time, disadvantaged families typically pay lower childcare fees in some regions, which might make them less responsive to the free daycare policy. In the end, it is an empirical question whether the policy benefits some families more than others.

To test for heterogeneity in treatment effects across population subgroups we use our baseline model in equation (2.2) but allow the coefficients on the treatment variable to vary for population subgroups defined by education and marital status of the parent or household income.[32] We estimate the model separately for each group.[33]

Table 2.7 shows two interesting patterns: first, low-skilled and poor parents respond much more to the free daycare policy than skilled or richer parents (see the first and second panel in Table 2.7). Public daycare for children between the ages of 2 and 6 increases by 5.6 percentage points for low-skilled and even 8.1 percentage points for poor children compared to no change in the average family – an increase by 8% (0.056/0.682) and 12% (0.081/0.701) respectively. Mirroring the sharp increase in daycare attendance, there is a sizable decline in exclusive home care by 7.9 percentage points in low-skilled families and

[32] We use the official definition of poverty in Germany. Accordingly, households are classified as poor if they have an income (adjusted for size) below 60% of the median household income.

[33] Alternatively, one could also estimate the model jointly to account for potential correlations between the socio-demographic variables. The results (not reported) are actually very similar, which is explained by the low correlation (at most 0.28) among the three demographic characteristics.

Table 2.6: Eligibility for Free Childcare and Child Outcomes

	2-3 Year-old Children Cognitive/Noncognitive Skills (Vineland Scale) (1)		5-6 Year-old Children Behavioral Problems (SDQ Score) (2)
Vineland Adaptive Behavior Scale	-0.017	Strengths and Difficulties Questionnaire	0.005
	[0.036]	(SDQ Score)	[0.073]
Motor Skills	-0.007	Conduct Problems	-0.079
	[0.032]		[0.065]
Skills in Daily Activities	-0.146***	Emotional Problems	-0.034
	[0.026]		[0.122]
Language Skills	-0.002	Problems with Peers	0.010
	[0.026]		[0.099]
Social Skills	0.083	Attention Problems	0.063
	[0.081]		[0.082]
Observations	5,488	Observations	2,386

Notes: The dependent variables in column (1) are child outcomes of 2-3 year-old children living in West Germany between 2005 and 2016. The data on non-cognitive skills come from the supplementary "mother-child" and the "your child between age 2 and 3" questionnaires, which ask additional questions of mothers with children born in 2003 or later. Mothers report whether a child is not able (=0), partly able (=1) or fully able (=2) to perform a certain skill. The adapted Vineland Maturity Scale consists of 20 items in total where each of its four subcategories (social skills, motor skills, daily activities, language skills) contains 5 questions. All scores are standardized to mean 0 and standard deviation of 1 in our sample. Larger scores mean that a child is better able to perform the specified skill. The table reports the coefficients on the eligibility indicator, which is defined as in earlier tables (in terms of a standard deviation). The dependent variables in column (2) are child outcomes of 5-6 year-old children living in West Germany between 2008 and 2016. The data on non-cognitive skills come from the supplementary questionnaire answered by mothers which elicits a version of the Strengths and Difficulties Questionnaire (SDQ) suggested by Goodman (1997). Mothers answer "not true", "somewhat true" and "certainly true" to 17 statements on socio-emotional behavior over five separate dimensions: Emotional symptoms, conduct problems, hyperactivity/inattention and peer relationship problems. The dimensions are also summed to a Total Difficulties Score (SDQ Score). Each score is normalized to have mean zero and a standard deviation of one in our West German sample between 2008 and 2016. Larger scores mean that a child has more problems in the specific socio-emotional dimension (in terms of a standard deviation). All specifications include as controls: child characteristics (gender and detailed age), controls for the mother (age, marital status, foreign citizenship and education) and household characteristics (household size, the number of children and whether there is an infant under 1). We also include birth cohort, age group, state fixed effects, state x age groups and cohort x age group fixed effects as well as interview month fixed effects. All standard errors are clustered at the state level. * p<0.1, ** p<0.05 and *** p<0.01.

Source: Socio-Economic Panel (2005-2016) for column (1); and Socio-Economic Panel (2008-2016) for column (2).

10.6 percentage points in poor families. Like the average households, low-skilled parents also buy more informal care. Children from low-skilled parents benefit a lot from access to free daycare in terms of their short-run cognitive and non-cognitive skills, while there is no effect for children from poor households. Single mothers respond like the average family in the sample: they use some of the additional funds to buy more informal childcare, which reduces the measured cognitive and non-cognitive skills of their child in the short-run. Also, there are few effects on the labor supply of single mothers.
Overall then, vulnerable families with preschool children are typically more responsive to universally offered free daycare policies than the average family. For low-skilled and poor households, the policy closes a sizable fraction of the attendance gap: in the pre-policy period, children from low-skilled parents are 13.5 percentage points less likely to attend daycare; introducing a free daycare slot reduces this gap by 40%. For children from poor households, the pre-policy attendance gap is 13.2 percentage points is closed by almost two-thirds.

Table 2.7: Heterogeneity of Effects for Population Subgroups

	Public Daycare Ages 2-6 (1)	Informal Childcare Ages 2-6 (2)	Childcare at Home Ages 2-6 (3)	Female LFP Ages 2-6 (4)	Child Outcomes Ages 2-3 (5)	Child Outcomes Ages 5-6 (6)
Eligible Child	0.002	0.050**	0.020	0.014	-0.123***	0.017
	[0.018]	[0.017]	[0.018]	[0.019]	[0.023]	[0.080]
Eligible Child*Low-skilled HH	0.056***	0.001	-0.079***	-0.051	0.245***	-0.085
	[0.014]	[0.027]	[0.018]	[0.032]	[0.025]	[0.150]
Eligible Child	0.002	0.030	-0.001	0.015	-0.084***	0.013
	[0.017]	[0.020]	[0.014]	[0.020]	[0.020]	[0.077]
Eligible Child*Poor Household	0.081***	0.054	-0.106***	-0.038	0.084**	-0.053
	[0.012]	[0.037]	[0.011]	[0.025]	[0.031]	[0.136]
Eligible Child	0.015	0.036	-0.018	0.015	-0.076***	0.016
	[0.016]	[0.020]	[0.013]	[0.017]	[0.021]	[0.077]
Eligible Child*Single Parent	-0.017	0.023	0.010	-0.034	0.010	-0.066
	[0.018]	[0.049]	[0.013]	[0.023]	[0.049]	[0.120]

Notes: The dependent variables in columns (1) to (3) are childcare choices of children aged between 2 and 6 in West Germany between 2000 and 2016; the dependent variable in column (4) is female labor force participation. The dependent variable in column (5) is the Vineland Adaptive Behavior Scale for 2-4 year-old children; in column (6), it is the total score from the Strengths and Difficulties Questionnaire. In each of the three panels, the table reports the coefficient on eligibility for free daycare and its interaction with the population subgroup specified (low-educated, poor or single parents). Low-educated parents have not completed a high school degree or vocational training. Poor parents have a household income of less than 60% of the median household income in Germany (adjusted for household size), while single parents live in households with no other adult. All specifications include as controls: child gender and age group, the same parental and household characteristics as in previous tables, state unemployment and GDP per capita (linear and squared terms), state and birth cohort fixed effects as well as state x age group and cohort x age group fixed effects. In columns (5) and (6), we also include detailed child age (in 3-month windows) and interview month fixed effects. Standard errors are clustered at the state level. *** p<0.01, ** p<0.05 and * p<0.1.

Source: Socio-Economic Panel (2000-2016).

2.6 Robustness Analysis and Standard Errors

Section 2.6.1 reports placebo tests to check for differential shocks prior to the reform and a range of specification checks to demonstrate the robustness of our findings. Section 2.6.2 discusses selective migration, while Section 2.6.3 reports alternative ways of estimating standard errors.

2.6.1 Placebo and Other Specification Checks

The graphical evidence in Figure 2.1 (in Section 2.3) suggests no differential pre-trends. We next investigate more systematically whether eligible families experience any differential shocks in the pre-reform period compared to children of similar ages and birth cohorts in states that have not (yet) adopted a free daycare policy. To do so, we perform placebos by shifting the free daycare policy in adopting states two, four and six years prior to the actual reforms. Table 2.8 shows the results: the top panel reports the mean effect; the bottom panel allows the treatment effect to vary by age group. Eleven of the 12 average effects are statistically indistinguishable from zero. Only one (female labor supply six years before an actual reform) is statistically significant as we would expect based on a significance level of 5-10%. Turning to the effects by age group in the bottom panel, only one out of 48 coefficients (informal care six years prior to an actual reform) is statistically different from zero. Overall then, the evidence on placebo reforms supports our identifying assumption that there were no differential trends between families living in treatment and control states.

Even if the placebo reforms do not reveal statistically significant deviations in the treatment states prior to a reform, other omitted factors could affect our estimates. We investigate the role of other influences in Table 2.9 for the average effect and in Table 2.12 stratified by child age. Our first robustness check removes all parental and household characteristics from our baseline to show that our treatment effects are unaffected by controlling for observables. The first row of Table 2.9 (and Table 2.12) shows that estimates obtained by including only state, birth cohort and age fixed effects as well as state x age and cohort x age interactions are very similar to the baseline. Hence, parental choices do not depend on the set of observable characteristics included as controls.

A second concern are that children belonging to the same birth cohort will differ in age across states for two reasons (see Section 2.3): first, the cutoff month used to define the child's birth cohort differs across states. Second, we observe children and their families at different points during the year as the survey is undertaken year-round (though 90% of the interviews are between January and August).[34] Our second robustness check therefore

[34]As the school year lasts around eleven months, two children born on the same date may differ in age

includes controls for 3-month age windows (60-62 months, 63-65 months etc.) as well as interview month fixed effects. Accounting for detailed age and survey month yields estimates (shown in the second row of Table 2.9) very similar to the baseline.

An alternative approach to control for age effects is to use a narrow sample of children born around the cutoff date for school entry. In the third row, we re-estimate equation (2.2) but restrict the sample to children born up to 4 months before or after the cutoff month. The coefficients are again very similar to the baseline results suggesting that age differences cannot account for our results.

Shocks or differences other than age occurring around the reform date might affect the interpretation of our estimates. One such change is the expansion of daycare slots for children under the age of three as discussed in Section 2.2.1. If treatment states expand their daycare slots for young children around the same time as they introduce free daycare and to a larger extent than control states, our estimation strategy would identify the combined effect of the increase in slots supplied and lower daycare prices. To check whether changes on the supply side have an effect on our estimates, we re-estimate the baseline in equation (2.2) controlling for the supply of daycare slots per 100 children at the district level. The results reported in the fourth row of Table 2.9 (and in Table 2.12) show that our estimates remain unchanged. Therefore, the increase in daycare attendance and decline in home care with no effects on female labor supply are indeed behavioral responses to the price decline and not a reaction to the availability of daycare. The only exception is that the increase in informal childcare is somewhat larger conditional on the supply of slots. Here, the baseline coefficient on informal care is likely a lower bound of the true effect.

A fourth concern is that we do not account for the cumulative nature of the policy in the treatment states. A child born in 2013 in *Hamburg*, for instance, has been eligible for free daycare since age 2. Hence, that 2-year-old may attend daycare free of charge for up to four years in *Hamburg*, while a child belonging to the same cohort in *Bremen* has no access to free daycare and would have access to just one free daycare year in *Bavaria*. The treatment variable then varies from zero years for children in non-adopting states up to four years in the states with the most comprehensive free daycare policy after the phase-in (see Table 2.1). We then use the cumulative number of years eligible for a free daycare slot as an alternative treatment variable in equation (2.2).[35] The fifth row in Table 2.9 (and Table 2.12) shows similar effects for informal care; now, we also find a

by up to 11 months depending on the date of the interview. The raw data suggest that treated households are interviewed somewhat later. A regression of the interview month on our baseline specification (child, parent and household demographics, state, birth cohort and broad age group fixed effects) yields no statistically significant relationship between treatment and interview month.

[35] Using the cumulative number of years of eligibility to free daycare as treatment variable thus accounts for the intertemporal decision-making of households.

statistically significant increase for daycare and a decrease in home care. Hence, offering multiple years of free daycare increases the impact of a free daycare policy on eligible families.

We further check whether there are any anticipation effects: families with children under the age of 5 might send their child to daycare earlier if that child becomes eligible for a free daycare slot in the last daycare year. To test for anticipation, we drop children in the last daycare year and all families in the states adopting reforms for younger children. Our results in the fifth row of Table 2.9 (and Table 2.12) suggest that parents indeed send their child to daycare earlier when their child will be eligible for free daycare in the future.

As the effects for younger children (below the age of 5) are identified from the three states that adopted comprehensive reforms only, we also check whether the effects differ for these treatment states. Such differences would cast doubt whether we can generalize the effects for younger children identified from these three states to the whole of West Germany or even other countries. The final row reports the effects in the comprehensive reform states: the increase on informal care and female labor supply is very similar to the estimates in the full sample (see Table 2.4 and 2.5). The coefficients on public daycare and exclusive home care are slightly stronger than in the baseline reported in Table 2.4, but none of the coefficients reaches statistical significance. It therefore seems plausible to assume that a free daycare policy has similar effects in all states if they introduced a free daycare policy for all preschool children.

2.6.2 Selective Migration of Eligible Families

Another concern with our estimation strategy might be that families with preschool children selectively migrate into states that adopt a free daycare policy. There is some anecdotal evidence that local governments indeed aim to attract families with young children by providing free daycare slots in addition to other benefits. If there is selective migration into adopting states, our estimates on childcare arrangements, for instance, might not reflect behavioral responses of eligible families, but rather a change in the mix of eligible households residing in reform states.

To assess this concern, we collected migration statistics by detailed age groups from the State Statistical Offices. We obtained comparable information for six of the nine reform states.[36] The net inflow of families with children under the age of 6 into adopting states increases after the reform. Controlling for state and school year fixed effects as well as local economic conditions, a free daycare slot in the last year attracts around 630

[36]Data are available for *Bavaria*, *Berlin*, *Hamburg*, *North-Rhine-Westphalia*, *Rhineland-Palatinate* and *Schleswig-Holstein*.

Table 2.8: Placebo Tests

	Public Daycare			Informal Childcare			Exclusive Care at Home			Female Labor Force Participation		
	(t-2)	(t-4)	(t-6)	(t-2)	(t-4)	(t-6)	(t-2)	(t-4)	(t-6)	(t-2)	(t-4)	(t-6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Placebo Free Daycare	0.013	0.019	0.031	-0.005	-0.032	0.037	-0.004	-0.003	-0.021	0.025	0.029	0.070**
	[0.014]	[0.021]	[0.027]	[0.015]	[0.023]	[0.027]	[0.008]	[0.016]	[0.017]	[0.017]	[0.016]	[0.024]
Placebo Free Daycare	0.032	-0.002	0.003	-0.051	-0.066	-0.019	0.003	0.028	-0.012	0.037	0.057	0.089
	[0.028]	[0.032]	[0.038]	[0.087]	[0.047]	[0.054]	[0.032]	[0.047]	[0.035]	[0.051]	[0.076]	[0.076]
Placebo*Ages 3-4	-0.024	0.021	0.021	0.086	0.006	0.102	-0.022	-0.050	-0.014	0.028	-0.012	-0.063
	[0.028]	[0.022]	[0.019]	[0.105]	[0.062]	[0.083]	[0.043]	[0.052]	[0.031]	[0.036]	[0.090]	[0.058]
Placebo*Ages 4-5	-0.036	0.018	0.017	0.086	0.011	0.128**	0.002	-0.034	0.009	-0.012	-0.062	-0.057
	[0.025]	[0.017]	[0.029]	[0.084]	[0.071]	[0.050]	[0.035]	[0.048]	[0.032]	[0.047]	[0.054]	[0.061]
Placebo*Ages 5-6	-0.029	0.026	0.046	0.020	-0.004	0.067	0.005	-0.023	-0.005	-0.051	-0.064	-0.011
	[0.037]	[0.030]	[0.040]	[0.093]	[0.050]	[0.060]	[0.036]	[0.049]	[0.040]	[0.053]	[0.087]	[0.078]

Notes: The dependent variables are the three childcare choices (in columns (1)-(9)) and female labor supply decisions (in columns (10)-(12)). In the top panel, the coefficients are from separate regressions of placebo reforms 2 years (first specification), 4 years (second specification) or 6 years (third specification) prior to the actual adoption of free daycare for a child in a certain age group in the state of residence. The bottom panel shows the same placebo indicator interacted with the age group of each child. All specifications include state, birth cohort, age group, state x age group and cohort x age group fixed effects as well as the eligibility indicator (plus its interaction with age group in the bottom panel). We further include the same aggregate state controls, parental and household controls as well as child gender in the baseline Tables 2.2 - 2.4. See notes to Tables 2.4 and 2.5 for details. Standard errors are clustered at the state level. *** p<0.01, ** p<0.05 and * p<0.1.

Source: Socio-Economic Panel (2000-2016).

Table 2.9: Specification Checks

	Public Daycare (1)	Informal Childcare (2)	Childcare at Home (3)	Female LFP (4)
No Demographic Controls	0.012	0.041	-0.019*	0.009
	[0.011]	[0.023]	[0.010]	[0.024]
Control for Child Age and Interview Month	0.004	0.039	-0.013	0.009
	[0.010]	[0.026]	[0.010]	[0.018]
Using Sample around Month of Birth Cutoff	0.015	0.041**	-0.026*	0.005
	[0.015]	[0.018]	[0.012]	[0.017]
Control for Supply of Daycare Slots	-0.010	0.067**	-0.019	-0.022
	[0.014]	[0.026]	[0.016]	[0.028]
Cumulate # Years Eligible	0.034***	0.011*	-0.021**	0.006
	[0.008]	[0.005]	[0.007]	[0.010]
Anticipation Effect	0.024*	0.009	-0.012	-0.025*
(2-5 year-olds, eligibile at ages 5-6)	[0.011]	[0.017]	[0.016]	[0.012]
Comprehensive Reform States Only	0.036	0.033	-0.037	0.003
(drop last year of daycare)	[0.046]	[0.016]	[0.041]	[0.040]

Notes: The table reports several specification checks for the three childcare choices and maternal labor supply shown in the top row. Each coefficient and standard error (in square brackets) come from a separate regression. The first row includes only state, birth cohort and age fixed effects as well as cohort x age group and state x age group interactions; all other specifications include the same controls as in Tables 2.4 and 2.5. See notes to Tables 2.4 and 2.5 for details. The second row adds controls for interview month and child age (in 3-month windows) to the baseline specification in equation (2); the third row restricts the sample to children who are born 4 months before or after the cutoff birth month for entry into a given daycare year. The fourth row controls for the supply of public daycare slots per 100 children for 0-3 year-olds and 3-6 year-olds in the family's district of residence. The fifth row uses as key independent variable the cumulative number of years a child is eligible for free daycare: up to 4 years for a 2-year-old child, up to 3 years for a 3-year-old child etc. The sixth row tests whether there is any anticipation effect, i.e. whether children aged between 2 and 5 change their childcare or labor supply choices when the child is eligible for the last year (at age 5-6). The final row restricts the analysis to children aged between 2 and 5 (and hence, to states who adopted free daycare for more than the last daycare year). Standard errors are clustered at the state level. *** $p<0.01$, ** $p<0.05$ and * $p<0.1$.

Source: Socio-Economic Panel (2000-2016).

additional families with preschool children. At the same time, however, migratory flows of families with young children are extremely low (only about 4-5%) relative to total in- and outflows across state borders over the 2000-2016 period. Furthermore, if we compare the number of net inflows to the stock of families with preschool children in each state using the 2011 Census, the total inflow of preschool children makes up only about 0.5-1% of the population in that age range. As such, we think that the somewhat higher inflows of preschool children following a free daycare reform does not pose a serious challenge to the interpretation of our findings.

2.6.3 Alternative Estimates of Standard Errors

Our main analysis clusters standard errors at the state level as the reforms were introduced by state governments, and all households with children in a certain age range were affected by the same reform. As our clustering strategy might be sensitive to the type of clustering as well as the small number of clusters (N=11), we report in Table 2.13 a range of alternative strategies for obtaining standard errors. As in our main Tables 2.4 and 2.5, we estimate variants of model in equation (2.2) where the dependent variables are again childcare choices and maternal labor force participation. The first specification (in odd columns) reports the average effect across all age groups, while the second specification (in even columns) shows interaction effects with child age.

Following Abadie et al. (2017), we first present estimates without clustering. The rationale is that fixed effects at the level of clustering (in our case, the state) will take care of correlated errors as long as there is no heterogeneity in treatment effects. Alternatively, we also explore the sensitivity of the estimated standard errors to clustering below the state level. Here, we cluster at the state-year level, given that policies evolved over time as well as across states. The second strategy includes separate state clusters for the pre- and post-policy period to allow for breaks in the temporal dependence of the error terms over time. For all three alternative estimators, the resulting standard errors are sometimes larger and sometimes smaller than in the baseline. Most importantly, our conclusion that a free daycare policy has increased informal childcare mostly for the youngest children (raising daycare attendance, informal childcare and female labor force participation, while reducing care at home) is supported by Table 2.13 irrespective of whether we do not cluster at all or cluster below the state level.

The estimated standard errors with clustering might still be sensitive to the small number of clusters, which is 11 in the case of clustering at the state level or 22 in the case of state x pre-/post-reform period. To address this concern, we implement a wild bootstrap procedure as proposed by Miller et al. (2008) and Cameron and Miller (2015). This procedure broadly generates confidence intervals and p values that support our main

results: a free daycare policy raises the use of informal childcare and shifts childcare arrangements for the youngest children and the labor supply of their mothers.[37] Overall then, our extensive set of robustness checks confirms that the conclusions presented in Section 2.5 remain qualitatively valid.

2.7 Conclusion

We investigate how the introduction of free public day care affects child care attendance, maternal labor supply and short-run child development. For estimation, we exploit quasi-experimental variation in childcare prices induced by the adoption of free daycare polices in nine out of eleven states in West Germany.

Our findings suggest that childcare attendance encourages earlier entry into public daycare for the youngest children (aged between 2 and 3), mirrored by a decline in exclusive care at home. Informal care also increases, suggesting that formal and informal daycare are complements in the German context. Reflecting the increase in daycare attendance among young children, mothers increase their labor supply, especially at the extensive margin. Hence, for the youngest children, a free daycare policy can encourage behavioral changes both among children and mothers. These behavioral changes have little effect on the average child in terms of their short- and medium-run development: while there seems a slight decline in daily skills, social skills actually improve.

For older children (aged between 3 and 6), we find little response in public daycare attendance, which had been high even before the policies were adopted. Mothers of older children aged between 4 and 6 mostly respond by increasing their labor supply at the intensive margin: they are more likely to work full-time and work longer hours. Given that the share working full-time and the average number of hours remain low in this group, a policy of offering free daycare seems not enough to obtain more full-time participation in the labor market among mothers in Germany – at least not in the short-run.

Finally, we also find sizable heterogeneity in the estimated effects: poorer and low-skilled households respond much more to the free daycare policy than the average household. Daycare attendance in this group increases much more than for the average child, which substantially reduces the pre-reform enrollment gap of around 13 percentage points between poor and rich or between low-skilled and skilled parents by up to two-thirds. It remains an open question, however, whether more targeted measures, which reduce the

[37] The wild bootstrap procedure uses binary weights for bootstrapping; we also implemented the Webb (6 points) weights, which have performed better in Monte Carlo simulations when there are less than 10 clusters (Cameron and Miller 2015). The Wald test statistics generated by this wild bootstrap procedure lead to conclusions that are very similar to those of the standard wild bootstrap procedure reported in Table 2.13.

fiscal burden of a universal policy of free daycare, could achieve a similar outcome at lower cost.

2.A Appendix

2.A.1 German Socio-Economic Panel (2000-2016)

The GSOEP provides comprehensive data on a representative sample of German households. We focus on households with young children and study the following variables:
Childcare variables: Our main dependent variables are the type of educational institution (school, kindergarten or other daycare facility) each child until the age of 6 currently attends if at all. Based on this information, we code whether a child attends a public childcare facility or not. We denote all childcare facilities that are publicly subsidized as public daycare; publicly subsidized childcare may be provided by the local community, churches, companies or other non-profit organizations. If the child attends an educational institution, the parents are asked whether the child attends only in the morning, only in the afternoon or the whole day. The survey also inquires about regular childcare provided by persons outside the household. These external providers could be relatives not living in the household, neighbors, friends or a paid child minder We define an indicator variable equal to one if any type of informal childcare is used. The variable is coded as zero if no informal childcare is used. In some specifications, we also distinguish whether the care is provided informally by a relative, friend or neighbor or whether it is purchased on the informal market from a child minder or nanny. Information about these informal sources of childcare is available for all years except 2003. Finally, we define the variable exclusive care at home as equal to one if no public or informal childcare outside the household is reported. Hence, home care does not necessarily imply that all childcare is provided by the parents, because it includes childcare by people living in the same household (like grandparents, au pairs or older siblings, for example). The variable is equal to zero if the child attends public childcare or is cared for by other people outside the household.
Maternal labor supply: We code labor force participation equal to one if the individual works full- or part-time, is marginally employed ("geringfügig beschäftigt"), is currently in school or vocational training. A mother is working full-time if she works 30 or more hours per week; working hours refer to the number of hours per week in the work contract.
Child outcomes: Data on child outcomes for 2-3 year-old children are taken from a supplementary questionnaire answered by mothers with children born in 2002 or later. The data are available annually since 2005. We use the questions on social, language and motor skills and skills for daily life to assess the short-run effects of the new policy on outcomes for eligible children. The skills elicited come from a version of the Vineland Adaptive Behavior Scale which has been adapted to the time constraints of a general household survey. *Social skills* cover the following tasks: whether the child calls familiar people by name; whether the child plays games with other children; whether the child participates in role playing games; whether the child shows liking for certain playmates; whether the child calls his/her own feelings by name. For *motor skills*, the survey asks to assess whether the child walks down the stairs forwards; whether the child uses door handle to open doors; whether the child climbs jungle gyms and other high playground equipment; whether the child uses scissors to cut paper and whether the child draws recognizable figures. For *language skills*, the following items are assessed: whether the child understands brief instructions; whether the child forms sentences with at least two words; whether the child speaks in full sentences of at least four words; whether the child

listens attentively to a story for at least 5 minutes; and whether the child can relate simple messages. Finally, the set of *skills in daily activities* comprises: whether the child eats with spoon without making a mess; whether the child blows his/her nose without assistance; whether the child uses the toilet "to do number two"; whether the child can put on pants and underpants correctly; and whether the child brushes teeth without assistance. For each question, the mother assesses the ability of her child on a 3-point scale: 1=yes, 2=to some extent and 3=no. From the individual items, we construct a score for the four categories by summing over the answers to each item coding as 0 if the child cannot perform the skill, as 1 if the child partially and as 2 if the child fully performs the skill. Each score ranges from a minimum of 0 to 10. We also calculate a total score as the unweighted sum over the four categories; the total score then ranges from 0 to 40. We then normalize the score to have zero mean and a standard deviation of one in our sample of 2-3 year-old children in West Germany from 2005-2016. A higher score means that the child is better able to perform a specific (set of) task(s).

To analyze the short-run effects on eligible children in older age groups we make use of a shorter version of the Strengths and Difficulties Questionnaire for 5-6 year-old children which has been available since 2008. The questionnaire asks: "Compared to other children of the same age how would you assess your own child?". Then, a list of 17 skills is presented. On a scale from 1 to 7 parents can choose whether their child is rather talkative or still, rather untidy or neat, good-natured or irritable, not interested or hungry for knowledge, has good confidence or is insecure, is withdrawn or outgoing, focused or dis-tractable, defiant or obedient, understands quickly or needs more time and is anxious or not. We construct a total score from these items by first recoding the answers using the original Goodman scale (does not apply, applies somewhat, applies fully) (Goodman 1997). We then calculate the unweighted sum over all items and several subcategories. Finally, we standardize the score to mean zero and a standard deviation of one in our sample of 5-6 year-old children in West Germany for 2008-2016. A higher score reflects more behavioral problems.

Control variables: As additional control variables, we use household characteristics like household size, the number of children and whether there is an infant under the age of one in the household. As a measure of household income, we use monthly disposable household income measured in euros (deflated to 2010 prices). The specific question asks about the total sum of all income sources of the household adjusted for taxes and other contributions ("verfügbares Haushaltseinkommen"). A household is considered poor if the household income (adjusted for size using OECD equivalence scales) is below 60% of the median household income, the official definition of poverty in Germany.

To control for characteristics of the parent (or caretaker), we also code the age, education, marital status and whether one parent holds a citizenship outside the European Union. For marital status, we distinguish three categories: single (never married), married or in a long-term partnership and divorced or widowed. Educational attainment is defined as the highest educational level achieved. We define a person as low-skilled if she has no vocational training and no high-school degree ("Abitur"). A person is defined as medium-skilled if the highest educational degree is vocational training or a high-school degree. Finally, the person is high-skilled if she has a tertiary degree from a university or technical college. Further, the observation is coded as foreign if the parent does not have German citizenship.

To merge the parental information to the child record, we define the relevant caretaker of the child in the household. The survey contains an identifier for the mother of each child; if the identifier and hence mother is missing, we select the father of the child; if both parents are absent in the household, we choose a female adult (presumably a relative or close friend). In our sample, in more than 99% of all cases the responsible parent is the mother or another female adult living in the household. Our main results consider females as primary caretakers.
Aggregate economic controls: To control for state-specific labor market shocks, we include the state unemployment rate defined as percentage of registered unemployed people to the total number of employed persons. To control for the broader economic situation in each state, we also include GDP per capita. Both variables are available from the Federal Statistical Office.

2.A.2 Are Parental Assessments of Child Outcomes Reliable?

Parental assessments, often the only source of information on skills of very young children, may suffer from systematic biases. Caregivers may be positively or negatively biased in their perception, may give socially desired answers, or may report some behavior only because they are asked in the survey (e.g. Schwarz 1999). Yet, external validation studies of parent-reported data indicate that they are informative about the skills they are intended to measure. There is also little evidence that any bias in parent-based reports is correlated with the socio-economic characteristics of parents (De Los Reyes and Kazdin 2005; Treutler and Epkins 2003). Furthermore, a recent validation of the VABS used in the SOEP showed that maternal assessments are highly correlated with scores on an examiner-administered test of infant development (Sandner and Jungmann 2016).
Maternal assessments of their child may also be affected by the time a mother spends with the child. Mothers might become less critical, for instance, as they care for their child at home and observe the child's eating habits or language use throughout the day. In that case, a change in maternal assessments might be the result of changes in childcare arrangements induced by the free childcare policy - and not the result of an actual change in the child's skill. In the absence of formal tests from developmental psychologists, we cannot address this concern directly. Yet, we can provide some indirect evidence that mere changes in perception are unlikely to drive our results. If maternal assessments mostly reflect the time spent with the child, they should not differ for children who attend formal or informal care (holding hours of care outside the home constant). Figure 2.3 shows regressions where the dependent variable is the total VABS score (in columns (1)-(3)) and the total SDQ score (in columns (4)-(6)). Key independent variables are the types of childcare used in addition to a number of socio-demographic characteristics. The results for children aged between 2 and 3 show that mothers assess their children more favorably if they attend public daycare instead of informal care (column (1)). The same pattern holds even if a child spends more time in formal than in informal care (column (2)). Finally, column (3) includes separate indicators for informal and formal care where the reference category is exclusive care at home. If there was a positive correlation between maternal assessments and care at home, we should see negative coefficients. Yet, we find the opposite pattern with maternal assessment being especially favorable if a child visits formal daycare.

The patterns are similar, but statistically weaker for children aged between 5 and 6 (shown in columns (4)-(6)). Recall that a higher SDQ score indicates more behavioral problems. Formal childcare is negatively correlated with behavioral problems (see columns (4) and (5)). The final column shows only a weak correlation between informal or professional care (relative to home care) and behavioral problems. One explanation for this weaker correlation is that only about 3% of the 5-6 year-old children in our sample are cared for exclusively at home, which leaves little variation in the data. Overall, the evidence in Table 2.11 indicates that maternal assessments do reflect more than biased perceptions of the mother.

2.A.3 Additional Tables

Table 2.10: Summary Statistics

	Pre-Policy Period (2000-2006)		Post-Policy Period (2007-2016)	
	Mean	Std. Dev.	Mean	Std. Dev.
Childcare Attendance	0.714	0.452	0.820	0.385
Full-time Attendance	0.207	0.405	0.269	0.444
Informal Childcare	0.402	0.490	0.315	0.465
Exclusive Care at Home	0.158	0.365	0.121	0.326
Maternal Employment	0.460	0.498	0.528	0.499
Full-time Work	0.177	0.382	0.183	0.386
Contractual Working Hours	20.39	10.42	21.63	10.28
Child is a Girl	0.493	0.500	0.484	0.500
Age of Child	4.38	1.24	4.26	1.29
Household Size	4.10	1.11	4.34	1.26
Number of Children	2.10	0.901	2.37	1.13
Infants under Age 1 in Household	0.025	0.157	0.056	0.231
Age of Mother	34.13	5.18	35.15	5.76
Mother Low-Skilled	0.193	0.395	0.203	0.402
Mother Medium-skilled	0.646	0.478	0.535	0.499
Mother High-skilled	0.144	0.351	0.228	0.419
Single Mother	0.068	0.252	0.110	0.313
Mother Married	0.882	0.323	0.842	0.365
Mother Divorced/Widowed	0.050	0.217	0.048	0.213
Foreign Mother	0.168	0.374	0.198	0.398
Unemployment Rate (%)	8.31	2.36	6.35	2.22
GDP per capita (Euros)	28754.7	4248.3	36207.0	5388.5
Observations	7,920		22,471	

Notes: The table reports summary statistics of our sample of preschool children (2-6 year-olds) and their parents in West Germany over the period from 2000 to 2016. The first two columns report summary statistics for the pre-policy period and the last two columns for the 2007-2016 period. Childcare arrangements are binary indicators equal to one if the family uses a certain childcare arrangement and zero otherwise. Low-skilled parents are those without a high school or vocational degree; medium-skilled parents have a high school or vocational degree and high-skilled parents a tertiary degree from university or technical college.

Source: Socio-Economic Panel (2000-2016).

Table 2.11: Are Maternal Assessments of Child Behavior Reliable?

	Cognitive and Noncognitive Skills (Vineland Adaptive Behavior Scale)			Behavioral Problems (Strengths and Difficulties Questionnaire)		
	(1)	(2)	(3)	(4)	(5)	(6)
Professional Childcare (relative to Informal Care)	0.073***			-0.101*		
	[0.019]			[0.055]		
Mainly Professional Care (more than Informal Care)		0.082***			-0.110*	
		[0.016]			[0.056]	
Informal Childcare			0.012			0.007
			[0.022]			[0.050]
Professional Childcare			0.102***			0.015
			[0.022]			[0.040]
Child is a Girl	0.202***	0.202***	0.188***	-0.290***	-0.296***	-0.293***
	[0.018]	[0.018]	[0.015]	[0.032]	[0.032]	[0.033]
Mother's Age	-0.005**	-0.005**	-0.005**	-0.015***	-0.015***	-0.017***
	[0.002]	[0.002]	[0.002]	[0.004]	[0.004]	[0.003]
Mother Medium-skilled	0.066**	0.068**	0.071**	-0.224***	-0.229***	-0.241***
	[0.026]	[0.027]	[0.023]	[0.060]	[0.066]	[0.072]
Mother High-skilled	0.106**	0.101**	0.110**	-0.436***	-0.447***	-0.426***
	[0.038]	[0.041]	[0.038]	[0.085]	[0.092]	[0.089]
Mother in School	0.077	0.074	0.056	-0.459***	-0.463***	-0.421**
	[0.047]	[0.047]	[0.044]	[0.117]	[0.121]	[0.153]
Mother Married	0.086***	0.080**	0.081***	-0.032	-0.014	-0.067
	[0.026]	[0.027]	[0.022]	[0.059]	[0.064]	[0.058]
Mother Separate/Widowed	0.026	0.018	0.008	0.119	0.136	0.074
	[0.051]	[0.051]	[0.053]	[0.114]	[0.117]	[0.123]
Mother Foreign-born	-0.071**	-0.071**	-0.066***	0.100	0.103	0.077
	[0.023]	[0.024]	[0.013]	[0.076]	[0.078]	[0.080]
Household Size	-0.030	-0.031	-0.036	0.072	0.076	0.054
	[0.029]	[0.031]	[0.027]	[0.062]	[0.061]	[0.059]
Number of Children in HH	0.052	0.054	0.059*	-0.143*	-0.152*	-0.132*
	[0.034]	[0.035]	[0.029]	[0.072]	[0.070]	[0.069]
Newborn Child in Household	0.014	0.016	-0.014	-0.028	-0.022	0.014
	[0.030]	[0.033]	[0.029]	[0.070]	[0.073]	[0.108]

Notes: The sample in columns (1)-(3) are 2-3 year-old children whose mothers answered the supplementary questionnaire between 2005 and 2016 (N=5,488); the sample in columns (4)-(6) are 5-6 year-old children whose mothers answered the supplementary questionnaire between 2008 and 2016 (N=2,339). In columns (1)-(3), the dependent variable is a standardized score on the Vineland Adaptive Behavior Scale. See notes for table 6 for a more detailed description of the dependent variable and controls included. In columns (4)-(6), the dependent variable is the standardized score of the Strengths and Difficulties Questionnaire (SDQ). The main independent vatriables are: in columns (1) and (4), an indicator variable equal to one if a child attends public daycare and zero if the child attends informal care; in columns (2) and (5), an indicator variable equal to one if the child spends more hours in public daycare and zero if it spends more time in informal care. In columns (3) and (6), the indicator variable for informal care is equal to one if the child attends informal care and zero if the child is cared for at home or in public daycare (and likewise for the indicator for professional childcare). The control variables are the same as in Table 2.6. All standard errors are clustered at the state level. * p<0.1, ** p<0.05 and *** p<0.01.

Source: Socio-Economic Panel (2005-2016) for columns (1)-(3); Socio-Economic Panel (2008-2016) for columns (4)-(6).

Table 2.12: Specification Checks for Eligibility by Broad Age Groups

	No Demographic Controls (1)	Child Age, Interview Month (2)	Sample of Birth Cutoff (3)	Supply of Slots (4)	# Years Eligible (5)	Anticipation Effect (6)	Comprehensive Reform States (7)
Public Daycare							
Eligible	0.055**	0.038	0.067	0.054**	0.061***	0.083***	0.130*
	[0.022]	[0.023]	[0.051]	[0.017]	[0.018]	[0.016]	[0.031]
Eligible*Ages 3-4	-0.051	-0.060**	-0.071	-0.071***	-0.083***	-0.096***	-0.146**
	[0.033]	[0.026]	[0.044]	[0.021]	[0.021]	[0.027]	[0.029]
Eligible*Ages 4-5	-0.031	-0.019	-0.087	-0.086***	-0.101***	-0.113**	-0.181
	[0.029]	[0.026]	[0.055]	[0.025]	[0.027]	[0.038]	[0.085]
Eligible*Ages 5-6	-0.066**	-0.045	-0.046	-0.084**	-0.099**		
	[0.025]	[0.027]	[0.062]	[0.027]	[0.032]		
Informal Childcare							
Eligible	0.086**	0.087***	0.078**	0.016*	0.010	0.006	0.038
	[0.033]	[0.012]	[0.026]	[0.008]	[0.006]	[0.019]	[0.019]
Eligible*Ages 3-4	0.012	0.002	0.020	0.005	0.010	0.001	0.023
	[0.024]	[0.026]	[0.024]	[0.011]	[0.010]	[0.012]	[0.044]
Eligible*Ages 4-5	-0.003	-0.009	-0.020	0.009	0.014	0.010	-0.044
	[0.021]	[0.016]	[0.019]	[0.013]	[0.009]	[0.009]	[0.024]
Eligible*Ages 5-6	-0.112**	-0.109***	-0.100**	0.050**	0.026*		
	[0.043]	[0.027]	[0.040]	[0.019]	[0.013]		
Care at Home							
Eligible	-0.079**	-0.065**	-0.095**	-0.029**	-0.031**	-0.020	-0.079
	[0.027]	[0.023]	[0.039]	[0.012]	[0.011]	[0.022]	[0.034]
Eligible*Ages 3-4	0.069***	0.079***	0.094**	0.028*	0.031**	0.012	0.067
	[0.018]	[0.020]	[0.031]	[0.014]	[0.013]	[0.018]	[0.032]
Eligible*Ages 4-5	0.049*	0.040*	0.097**	0.030*	0.035*	0.017	0.079
	[0.025]	[0.020]	[0.038]	[0.016]	[0.016]	[0.024]	[0.067]
Eligible*Ages 5-6	0.089**	0.073**	0.080	0.022	0.033		
	[0.030]	[0.026]	[0.047]	[0.019]	[0.021]		
Female LFP							
Eligible	0.088***	0.074***	0.074***	0.014*	0.012	-0.018	0.039
	[0.011]	[0.015]	[0.017]	[0.007]	[0.008]	[0.013]	[0.064]
Eligible*Ages 3-4	-0.040	-0.054	-0.054	-0.011	-0.013**	-0.011*	-0.013
	[0.025]	[0.039]	[0.038]	[0.009]	[0.006]	[0.006]	[0.097]
Eligible*Ages 4-5	-0.120**	-0.090**	-0.157***	-0.024***	-0.026***	-0.012	-0.115**
	[0.046]	[0.037]	[0.036]	[0.006]	[0.006]	[0.016]	[0.013]
Eligible*Ages 5-6	-0.118***	-0.092*	-0.080*	-0.048***	-0.026		
	[0.037]	[0.042]	[0.041]	[0.015]	[0.016]		

Notes: The table reports the same specification checks as in Table 2.12 but interacts the eligibility indicator with the broad age group of the child. The dependent variables are again the three childcare choices and maternal labor supply shown in the first column. All specifications include the same controls as in Tables 2.4 and 2.5. See notes to Tables 2.4 and 2.5 for details. Column (1) adds controls for interview month and child age (in 3-month windows); column (2) restricts the sample to children who are born 4 month before or after the cutoff birth month for entry into a given daycare year. Column (3) controls for the supply of public daycare slots per 100 children for 0-3 year-olds and 3-6 year-olds in the family's district of residence. Column (4) uses as key independent variable the cumulative number of years a child is eligible for free daycare: up to 4 years for a 2-year-old child, up to 3 years for a 3-year-old child etc. Column (5) tests whether there is any anticipation effect, i.e. whether mothers of children aged between 2 and 5 change childcare or labor supply choices when the child is eligible for the last year (at age 5-6). Column (6) restricts the analysis to children aged between 2 and 5 (and hence, to states who adopted a comprehensive free daycare policy). Standard errors are clustered at the state level. *** p<0.01, ** p<0.05 and * p<0.1.

Source: Socio-Economic Panel (2000-2016).

Table 2.13: Alternative Estimators for Variance-Covariance Matrix

	Public Daycare		Informal Childcare		Childcare at Home		Female LFP	
	Overall	by Age Group	Overall	by Age Group	Overall	by Age Group	Overall	by Age Group
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
No Clustering								
Eligible	0.013	0.063*	0.038**	0.083**	-0.016*	-0.079**	0.010	0.077**
	[0.010]	[0.037]	[0.017]	[0.039]	[0.009]	[0.034]	[0.016]	[0.036]
Eligible*Ages 3-4		-0.086**		0.002		0.099**		-0.057
		[0.042]		[0.056]		[0.039]		[0.052]
Eligible*Ages 4-5		-0.061		-0.003		0.065*		-0.100*
		[0.040]		[0.055]		[0.037]		[0.052]
Eligible*Ages 5-6		-0.053		-0.104**		0.077**		-0.092**
		[0.038]		[0.047]		[0.036]		[0.044]
State and Year Cluster								
Eligible	0.013	0.063	0.038**	0.083**	-0.016	-0.079*	0.010	0.077**
	[0.013]	[0.043]	[0.016]	[0.034]	[0.010]	[0.043]	[0.015]	[0.032]
Eligible*Ages 3-4		-0.086*		0.002		0.099**		-0.057
		[0.047]		[0.063]		[0.049]		[0.050]
Eligible*Ages 4-5		-0.061		-0.003		0.065		-0.100**
		[0.042]		[0.057]		[0.045]		[0.046]
Eligible*Ages 5-6		-0.053		-0.104**		0.077*		-0.092**
		[0.045]		[0.041]		[0.045]		[0.039]
State and Pre-/Post Policy Cluster								
Eligible	0.013	0.063*	0.038***	0.083*	-0.016	-0.079***	0.010	0.077***
	[0.015]	[0.031]	[0.011]	[0.043]	[0.011]	[0.026]	[0.015]	[0.019]
Eligible*Ages 3-4		-0.086***		0.002		0.099***		-0.057
		[0.025]		[0.052]		[0.025]		[0.038]
Eligible*Ages 4-5		-0.061**		-0.003		0.065**		-0.100***
		[0.026]		[0.061]		[0.024]		[0.030]
Eligible*Ages 5-6		-0.053		-0.104*		0.077**		-0.092**
		[0.039]		[0.051]		[0.032]		[0.035]
Wild Bootstrap								
Eligible	0.013	0.063	0.038	0.083	-0.016	-0.079	0.010	0.077
	[-0.018; 0.041]	[-0.004; 0.118]	[-0.006; 0.080]	[0.062; 0.103]	[-0.042; 0.009]	[-0.125; 0.024]	[-0.021; 0.042]	[0.045; 0.108]
	0.568	0.264	0.304	0.000	0.292	0.120	0.528	0.000
Eligible*Ages 3-4		-0.086		0.002		0.099		-0.057
		[-0.138; -0.023]		[-0.046; 0.051]		[0.048; 0.134]		[-0.126; 0.013]
		0.004		0.84		0.000		0.332
Eligible*Ages 4-5		-0.061		-0.003		0.065		-0.100
		[-0.106; -0.002]		[-0.027; 0.024]		[0.016; 0.104]		[-0.157; -0.031]
		0.04		0.964		0.008		0.116
Eligible*Ages 5-6		-0.053		-0.104		0.077		-0.092
		[-0.131; 0.037]		[-0.151; -0.058]		[0.010; 0.134]		[-0.167; -0.007]
		0.584		0.036		0.256		0.180

Notes: The table reports alternative approaches to clustering at the state level (reported in all other tables) in order to account for dependencies in standard errors: the first specification does not cluster at all (see Abadie et al., 2017); the second specification clusters by state and year, the third by state and the period before and after the policy change. In addition, the final specification reports estimates, the 95% confidence interval and the p-value from a wild bootstrap with 500 repetitions (see Cameron et al., 2008; Cameron and Miller, 2015). The dependent variables are childcare and female labor supply choices (shown in the top row) of families with preschool children in West Germany. The table shows the coefficients for the indicator whether a child is eligible for free daycare. All specifications include the same controls as in Tables 2.4 and 2.5. See notes to Tables 2.4 and 2.5 for details. *** p<0.01, ** p<0.05 and * p<0.1.

Source: Socio-Economic Panel (2000-2016).

2.A.4 Additional Figures

Figure 2.2: Provision of Public Daycare Slots

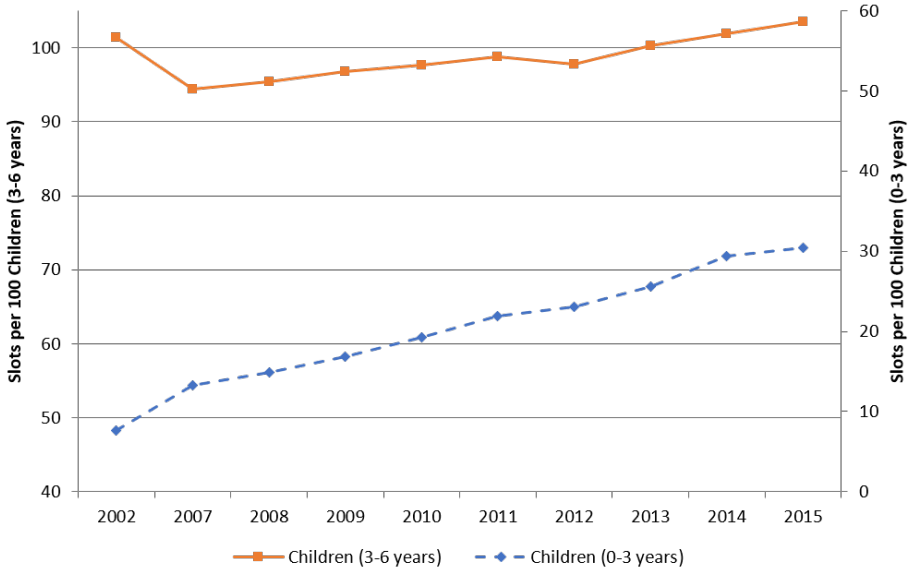
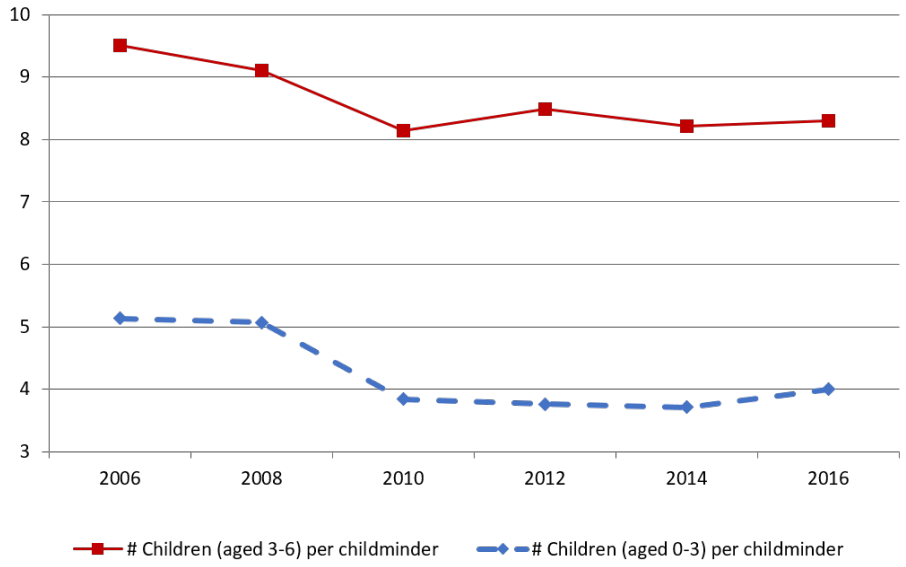


Figure 2.3: Evolution of Proxy for Childcare Quality



CHAPTER 3

Marginal Returns to Citizenship and Skill Development

joint work with Christina Gathmann and Christina Vonnahme

3.1 Introduction

Access to citizenship is the most fundamental integration policy a host country can offer its immigrant population. Naturalization has obvious benefits in terms of political rights and participation; but are there also benefits beyond the political realm? Some think that the importance of citizenship for economic or social integration has declined as other rights like permanent residency, for instance, have expanded over time.[1] Others, in contrast, argue that citizenship acts as an important catalyst for economic, social and even cultural integration.

Proponents of the view that citizenship matters for educational and labor market performance rely on three arguments: first, it improves the returns to labor market skills. Citizenship is not only a prerequisite for certain public sector jobs and some well-paid professions; it also removes any restrictions on career mobility that immigrants frequently face enabling them to work in any job, at any time and place.[2] In addition, employers might be more willing to invest in training an employee whose naturalization signals a long-term commitment to remain in the country (Lalonde and Topel 1997). To the extent that further career options and training offer better pay or working conditions than jobs available to the unnaturalized immigrant, citizenship improves the returns to labor market

[1] Shachar et al. (2017) provide a recent overview of the debate from the perspective of multiple disciplines.

[2] In some countries like Germany these restrictions apply to a much wider range of occupations: prior to 2012, non-EU citizens had only restricted access to regulated professions like lawyers, notaries, pharmacists or physicians. Also, non-EU citizens could not easily change their occupation prior to 2005 or move to a different EU member state without citizenship.

skills. Second, the offer to naturalize signals to immigrants that they fully belong to the host society, which might reduce return intentions and encourage further investments to succeed in the host country (Dustmann 2008). Finally, citizenship might also change the attitude of natives and immigrants alike reducing potential biases or overt discrimination, for instance (Felfe et al. 2018).[3]

Identifying how citizenship affects immigrant choices and immigrants' integration has enormous policy relevance. Migrant populations have grown rapidly in recent years, especially families with young children. Results from the Programme for International Student Assessment (PISA) indicate that between 2003 and 2015 alone, the share of students with direct migration experience or a parent who had migrated across international borders grew by six percentage points. By 2015, almost one in four 15-year-old students in the Organisation for Economic Co-operation and Development (OECD) were either foreign-born or had at least one foreign-born parent (OECD 2018b). Moreover, several countries have debated, announced or implemented reforms of their citizenship policy recently. The U.S. government has been debating restricting access to birthright citizenship for children of undocumented or illegal parents. In Europe, the U.K. has announced to tighten the language and knowledge tests necessary for naturalization. Germany, in contrast, has substantially liberalized its citizenship policy in recent decades.

Despite enormous policy relevance, there is little empirical evidence on the effects of access to citizenship. Identification is hampered by sizable empirical challenges. Countries with liberal citizenship policies, like citizenship by birthplace or short residency requirements for naturalization, typically differ along many other dimensions from countries with more restrictive access to citizenship. These cross-country differences are likely to influence immigrant selection, the environment immigrants find themselves in and hence, the choices families make after arrival. Access to citizenship sometimes varies within a given country: children with one native and one foreign-born parent obtain citizenship automatically, for instance. Yet, the selection into intermarriage and family circumstances are likely to differ for immigrant families with two foreign-born parents or with one foreign-born and one native parent. Therefore, using these within-country or cross-country differences is unlikely to identify causal effects of citizenship.

In this paper, we estimate the causal effects of citizenship and explore their heterogeneity for the skill development of children and young adults along both observable and unobservable characteristics. We focus in our analysis on young immigrants as immigrant-native gaps in language, math and science tests are sizable and persistent as has been demonstrated e.g. for 15-year-olds (Dustmann and Glitz 2011; OECD 2018b). Similarly,

[3]Note that this might reduce but not overcome discrimination as discrimination could be based on appearance or foreign sounding names as well.

the ethnic gap in achievement test scores and other skills increases substantially during childhood (Fryer Jr. and Levitt 2004; Heckman et al. 2006; Cunha and Heckman 2007). Given the strong link between student competencies and outcomes later in life, these gaps are likely to have long-term effects on the occupational careers and labor market success of children from immigrant families, reducing upward mobility and cementing unequal opportunities (Dustmann and Glitz 2011). Yet, immigrant-achievement gaps vary a lot across countries and are typically lower for traditional immigration countries like Canada or Australia than countries with restrictive citizenship policies like Germany or Sweden (Entorf and Minoiu 2005; Sweetman and van Ours 2015). This comparison suggests not only that citizenship policy could be important but also that effects are likely to be heterogeneous depending on the source country and socio-economic background of immigrant children.

Our setting provides a unique opportunity to identify citizenship effects and explore their heterogeneity across immigrant children. Over the recent decade, Germany has experienced a rapid increase of inflow of immigrants, increasingly families with children. While 23% of the overall population have a migration background, i.e. they or at least one parent was not born with German citizenship, more than one out of three children under the age of six now fall into this category (OECD 2018a). Traditionally, citizenship was tied to German descent. As such, naturalizations were rare and up to the discretion of the authorities. Two national reforms in 1991 and 2000 completely overhauled Germany's citizenship law. The 1991 reform introduced explicit criteria for how first-generation immigrants could naturalize. Immigrants who have reached the age of sixteen can naturalize if they have lived legally in Germany for at least eight years. Adults had to wait between eight and fifteen years but could include their dependent children in the application for citizenship. The second reform in 2000 introduced birthright citizenship for second-generation immigrants into German law. Children born in Germany after January 1, 2000 whose foreign-born parents had lived in the country legally for at least eight years are eligible for a German passport. A transitional rule allowed parents with at least eight years of residency to apply in 2000 to naturalize their child born in Germany between 1990 and 1999.

The 1991 and 2000 reforms together define four roads of eligibility to citizenship for children of immigrants: birthright citizenship, the transitional rule for immigrant children born in Germany, individual eligibility (at age 16) and eligibility through parents. All four categories of eligibility depend only on socio-economic characteristics like year of birth and year of arrival. They are, however, independent of any individual motivation and aspiration that are likely to affect both the decision to naturalize and possibly educational outcomes. As such, we use the four access options to citizenship as instruments

for the actual decision to naturalize and trace out its consequences for children's skill development.

To analyze who benefits from host country citizenship, we use the marginal treatment effect (MTE) framework introduced by Björklund and Moffitt (1987) and generalized by Heckman and Vytlacil (1999, 2005, 2007b), which relates the heterogeneity in the treatment effect to observed and unobserved heterogeneity in the naturalization decision. Previous studies on the determinants of naturalization decisions suggest that first-generation immigrant adults are typically positively selected with respect to formal education (Chiswick and Miller 2008; Gathmann and Keller 2018). If returns are heterogeneous, the marginal immigrant who gets naturalized under a more liberal citizenship policy might have zero or even negative returns.

Indeed, we find substantial heterogeneity in the returns to citizenship along both observable and unobservable characteristics. Immigrant children born in Germany and girls in particular benefit substantially from German citizenship in terms of language skills as measured by standardized test scores but also their school grade in German. At the same time, German-born children of immigrants are more likely to naturalize pointing to a positive selection on gains. Immigrant girls in turn, are just slightly more likely to obtain citizenship than immigrant boys. The positive selection in gains is reflected in unobserved heterogeneity as well: returns to citizenship with respect to language skills are declining with increasing resistance to treatment. Hence, children whose unobserved characteristics make them most likely to naturalize benefit the most, while children whose unobservable characteristics make them least likely to pick up German citizenship have zero returns to host country citizenship. We further show that improvements in language skills also help immigrant children to improve their school performance: children are much less likely to repeat a grade in school, for instance.

We then ask how potential reforms to citizenship policies would affect the skill development of (first- or second-generation) immigrant children. Given that we have limited common support for the propensity score in the MTE framework, we calculate marginal policy-relevant treatment effects, which analyze how the expansion or restriction of the take-up of citizenship would affect the skill development of immigrant children. Given the high take-up rates for immigrant children (around 80%) and the positive selection in gains, it comes at no surprise that expanding take-up even further yields few additional gains in language skills. Quite the contrary: returns to host country citizenship would actually be more pronounced if we restricted take-up to children with unobservable characteristics that make them most prone to naturalize ("low resistance" children). A possibly more attractive policy alternative to limiting access to citizenship is to expand access to birthright citizenship. At present, second-generation immigrant children only obtain

citizenship if one of their parents has lived in Germany for at least eight years or they become eligible when they turn sixteen. Traditional immigrant countries like Canada or the United States grant birthright citizenship to all children born in the country. Granting citizenship to all immigrant children born in Germany would carry substantial benefits in terms of language skills. The reason for these returns is that immigrant children born in Germany benefit the most from obtaining host country citizenship.

Our analysis makes several important contributions to the literature. Existing studies on the consequences of citizenship have estimated at most intention-to-treat effects of birthright citizenship on early childcare attendance, school entry and school track (Felfe et al. 2019). The study based on administrative data from a single German state finds very large effects, which closes most of the immigrant-native gap. Our analysis estimates causal effects (LATEs) of citizenship on skill development throughout primary and secondary school. Even more importantly, we explore for the first time the heterogeneity of returns to citizenship across observable and unobservable characteristics by estimating Marginal Treatment Effects. Furthermore, our data cover all states in Germany and allows to control for detailed parental characteristics like years since migration, which is typically not available in administrative datasets.

A second line of research analyzes how birthright citizenship has affected parental integration efforts (Avitabile et al. 2013), fertility behavior (Avitabile et al. 2014) and return migration (Sajons 2016) showing that parents increase their efforts to use the local language and reduce the total number of children in favor of more investments in the 'quality' of children. Another line of research investigates the consequences of access to citizenship on the labor market performance and marriage behavior of adult immigrants using panel data approaches (e.g. Bratsberg et al. 2002; Steinhardt 2012) or intention-to-treat effects (Gathmann and Keller 2018; Gathmann et al. 2019). Unlike all previous evidence, we focus on how access to citizenship through parents, individual eligibility or birthplace affects children's skill development, which is a crucial prerequisite for later success in the labor market. In addition, we are able to identify causal effects and explore for the first time their distribution along observable and unobservable dimensions.

Our study also contributes to the growing literature that estimates marginal treatment effects. So far, most studies focus on monetary returns to a college education (see e.g. Carneiro et al. 2011; Kaufmann 2014; Nybom 2017; Kamhöfer et al. 2018), to secondary education: (e.g. Carneiro et al. 2017: in Indonesia) or to early childcare education (Cornelissen et al. 2018; Felfe and Lalive 2018; Kline and Walters 2016).

This chapter proceeds as follows. We next discuss the background on the citizenship reforms, which introduced multiple paths of eligibility for citizenship. Section 3 then introduces the data sources and the policy variation in our data. Section 4 sets out the

econometric framework to estimate marginal treatment effects and discusses the estimation strategy. In Section 5, we present our main results on selection into citizenship and the returns to citizenship on skill development. Section 6 investigates how alternative citizenship policies might affect skill development by estimating marginal policy-relevant treatment effects. Finally, Section 7 concludes.

3.2 Background on Germany's Citizenship Reforms

For a long time, German citizenship was closely tied to ancestry. As such, there were few possibilities to naturalize unless one could demonstrate German descent. This traditional notion of citizenship was overhauled with the two citizenship reforms in 1991 and 2000.[4] The Alien Act (Ausländergesetz (AuslG)), which came into effect on January 1, 1991, defined for the first time explicit criteria how first-generation immigrants could naturalize in Germany. Most importantly, immigrants had to satisfy certain residency requirements, which were age-dependent: immigrants who came to Germany under the age of 8 or were born in Germany, could naturalize when they turned 16. An immigrant who arrived in Germany between the ages of 8 and 14 could naturalize after 8 years in Germany, while individuals arriving after age 14 required 15 years of legal residency.[5]

Applicants had to fulfill several additional criteria: first, they had to renounce their previous citizenship upon naturalization as the new law did not allow dual citizenship. Citizens of the European Union were exempt from this rule and could keep their original citizenship.[6] Second, the applicant must not be convicted of a severe criminal offense.[7] Furthermore, immigrants who arrived in Germany at age 15 or older had to demonstrate economic self-sufficiency, i.e. they should be able to support themselves and their dependents without welfare benefits or unemployment assistance. Younger immigrants who arrived in Germany before the age of 15 had to have completed a minimum of six years

[4] The citizenship reforms were brought on the agenda by the fall of the Iron Curtain and demographic changes on the one hand and political and court decisions on the other hand. Gathmann and Keller (2018) discuss the adoption process of the 1991 reform and Worbs (2014) provides an in-depth discussion of the 2000 reform.

[5] See AuslG (Alien Act). If the applicant stayed abroad for no more than 6 months, the period of absence still counted toward the residency requirement. Longer stays abroad (between 6 months and 1 year) may still count for the residency requirement if they are shown to be temporary.

[6] Children of bi-national marriages, for example, did not have to give up their dual citizenship until they turned 18. Exceptions were also granted if the country of origin prohibits the renunciation of citizenship or delayed it for reasons outside the power of the applicant, if the applicant was an acknowledged refugee or if the renunciation imposed special hardships on older applicants.

[7] Applicants with minor convictions, such as a suspended prison sentence up to 6 months (which would be abated at the end of the probation period), a fine not exceeding 180 days of income (calculated according to the net personal income of the individual), or corrective methods imposed by juvenile courts were still eligible. Convictions exceeding these limits were considered on a case-by-case basis by the authorities.

of schooling in Germany, of which at least four years had to be general education. It is important to stress that both requirements are less restrictive than those for obtaining a permanent work or residence permit. Finally, an applicant needed to declare her loyalty to the democratic principles of the German constitution. Foreign-born parents who fulfilled these requirements could include their dependent children (under the age of 18) in their citizenship application even if the child or spouse did not yet fulfill the individual residency requirement.

The Citizenship Act (Staatsangehörigkeitsgesetz (StAG)), which came into effect on January 1, 2000, introduced for the first time birthright citizenship into German law. Children born in Germany in 2000 or later obtain German citizenship automatically if at least one foreign-born parent has lived in Germany legally for 8 years (and had a permanent residency permit, available after five years of legal residence, for at least three years).[8] The law also stipulated a transitional rule for children born between 1990 and 1999: immigrant parents who had lived in Germany for at least eight years could apply for citizenship for their child. This option expired on December 31, 2000, however.

Furthermore, the 2000 reform reduced the residency requirements for immigrants arriving at the age of 15 or older to 8 years. Since 2000, most immigrants face residency requirements of eight years; only immigrants arriving under the age of eight have to wait until they turn 16 to become eligible. The other requirements of the 1991 reform stayed the same: applicants could not have a criminal record, had to demonstrate economic self-sufficiency and their loyalty to democratic principles. In addition, the new law also required applicants to demonstrate adequate German language skills prior to naturalization. As before, the law of 2000 did not recognize dual citizenship in general though exemptions became more common in practice.[9]

Together the reforms of 1991 and 2000 define four roads to obtaining citizenship for immigrant children: citizenship by birth, eligibility through the transitional rule, individual eligibility and eligibility through parents. Figure 3.1 provides an overview of the different access options and the associated requirements on residency, age and birthplace. For those with foreign citizenship, the 1991 reform defined two ways to naturalize: the child had to be at least 16 years old and lived in Germany legally for at least eight years (individual eligibility). Alternatively, a child under the age of 18 could obtain German citizenship if

[8] Eligibility is checked when parents register their newborn children, which is legally required in Germany. In most cases, the children also keep the citizenship of their parents' country of origin. Until 2014, children have to decide at age 21, which citizenship to keep ('option model'). Since December of 2014, children who have grown up in Germany for at least 8 years and have finished at least 6 years of formal education can keep both passports.

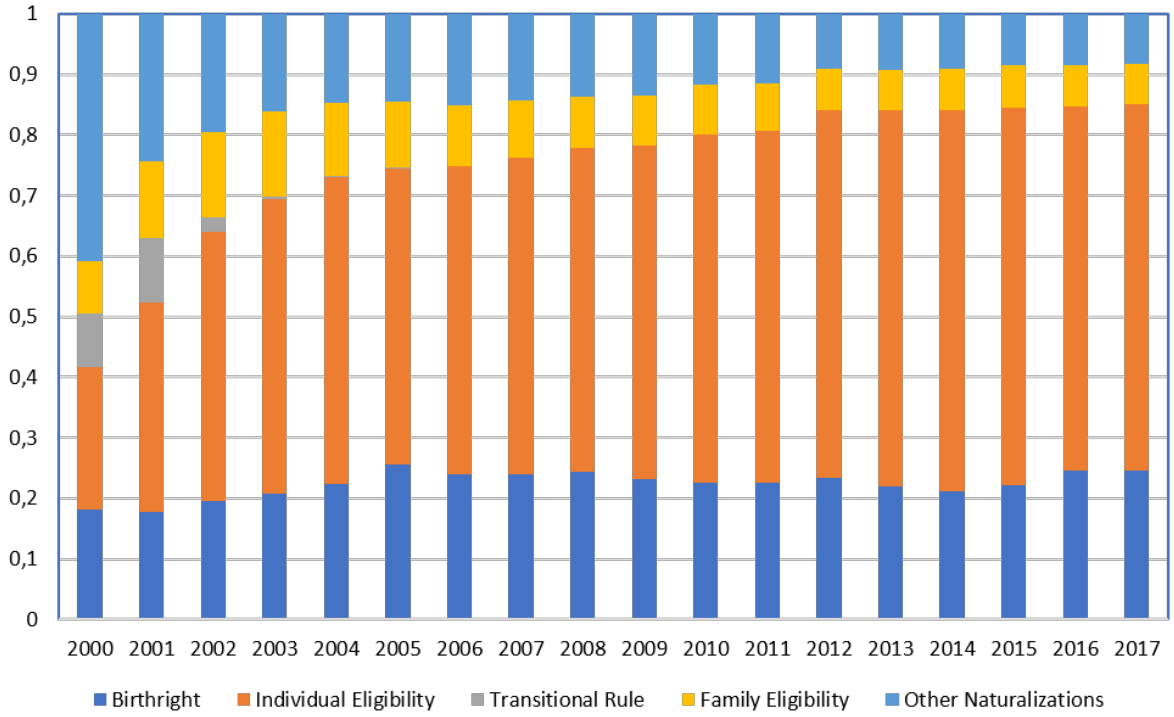
[9] It became easier for older applicants and refugees to keep their previous citizenship. Applicants could also keep their nationality if it was legally impossible to renounce it or if it imposed a special hardship like excessive costs or serious economic disadvantages (e.g. problems with inheritances or owning property in their country of origin).

Figure 3.1: Eligibility Rules for Citizenship

Sample with Foreign-born Parents	Child's Year of Birth: 1990-1999	Child's Year of Birth: 2000 or later
Child with Foreign Citizenship (Foreign- or German-born)	**Individual eligibility (1991 reform)** Child arrives aged 0-7: 9-16 years (age 16) Child arrives aged 8-14: 8 years (ages 16-22) **Eligibility through parent (1991/2000 reforms)** Parent arrives aged 8-14: 8 years Parent arrives aged 15 or older: 15 years [until 2000]; 8 years [since 2000]	
Child Born in Germany but with Foreign Citizenship	**Transitional rule** (2000 reform) if parent 8 years of legal residency in Germany	**Citizenship by birth** (2000 reform) if parent 8 years of legal residency in Germany

one foreign-born parent became eligible and included the child in the citizenship application (eligibility through parent). The 2000 reform further added two more options open to immigrant children born in Germany: for children born in 2000 or later whose foreign-born parents satisfy the residency requirement of 8 years (citizenship by birth) and for children born between 1990 and 1999 if their parents satisfied the residency requirement of 8 years and filed a citizenship application for their child in 2000 (transitional rule). Figure 3.2 shows the share of naturalizations granted in each category between 2000 and 2017. Across all age groups, the most important category is individual eligibility, which makes up for around 60% of all naturalizations over this period. Birthright citizenship has also played an important role accounting for around 20% of all naturalizations. The transitional rule, in contrast, played only a transitory role indeed: while 10% of all naturalizations in 2000 and 2001 occurred under the transitional rule, the fraction became negligible thereafter. Naturalization as part of a family application makes up less than 10% of naturalizations across all age groups. We next discuss the data we use to analyze the consequences of the two citizenship reforms on children's skill development.

Figure 3.2: Naturalizations by Type of Eligibility



Notes: The figure shows the percentages of naturalizations by birth under the transitional rule, individual eligibility, or as a family member of an eligible adult. Data prior to 2000 are not available as they did not distinguish between different legal forms of naturalizations.
Sources: Federal Office for Migration and Refugees (2008; 2019) and Federal Statistical Office (2018)

3.3 Data Sources

3.3.1 National Educational Panel Study

Our main data source is the National Educational Panel Study (NEPS, Blossfeld et al. 2011). The panel collects detailed information on children's education and skill development. The survey first samples schools and then randomly selects specific classes in the target grade. All students in the selected classes are asked to participate and then followed over time even if they repeat or skip a grade.[10] To select children under six who have not yet entered school, NEPS first samples primary schools and then surveys children in the day care centers that typically send children to the selected primary school. Our study covers the period from 2010 to 2017.

A unique feature of the data for our purpose is that the NEPS study administers standardized competence tests to all children. Such detailed standardized test results are rarely available in Germany and certainly not for large samples of school-aged children.[11] For our main analysis, we will focus on the evolution of language skills in German. Fluency in the host country language is crucial to learn many other subjects in school. In addition, tests of language skills are available for all children in several grades thus providing large enough samples for our analysis. Below, we will also investigate other outcomes like math, science or ICT skills as well as measures of school performance like grade retention or grades in German and math.

Similar to the PISA test program, the test for reading comprehension measures the ability to handle different types of texts encountered in daily life (like advertisements, literary or instructional texts).[12] The vocabulary test is based on the Peabody Picture Vocabulary Test and hence, similar to those in other large-scale panel studies (e.g. the British Cohort Study or the European Child Care and Education Study). Each test contains between 21 to 89 items and takes around 30 minutes to complete. For school children, the test is completed in the class room. Pre-school children or first-graders are tested one-to-one with an interviewer and age-adjusted items. We use the number of correct answers in each test as our baseline measure. To make effects comparable across subject and age groups, we rescale the scores to values between 0 and 100 and then standardize them to have mean zero and a standard deviation of one in our sample.

[10] The survey does not follow all students who leave the school, however. It is unclear a-priori whether these students are positively or negatively selected. We return to the question of selective attrition from the sample in more detail below.

[11] Prior studies on Germany have analyzed data on school readiness, which is assessed by a pediatrician when the child is five years old prior to school entry (Cornelissen et al. 2018; Felfe and Lalive 2018; Felfe et al. 2019).

[12] See the data appendix for a detailed discussion of all tests and measures of school performance analyzed.

A second unique feature of the NEPS data is that information on parents' education, working career and migration history is available. The survey interviews not only the child, but also the main caretaker, in 80% of the cases the mother, and collects information about the partner or spouse living in the household.[13] We restrict the sample to children where the main caretaker is foreign-born. The children might be born abroad or in Germany, i.e. are first- or second-generation immigrants. We keep children whose parents have naturalized at some point but exclude children with a native parent.

We focus our analysis on three cohorts spanning the period from kindergarten until the end of compulsory schooling. The first cohort, which we call the *child cohort*, is born in 2005 and 2006 and hence, on average eight years of age during our sample period. They are still in kindergarten in the first wave and go through primary school over our sample period. Most of this cohort is eligible for birthright citizenship, but could be eligible through their parents as well. The second or *teen cohort*, mostly born in 1999 and 2000, is on average 13 years of age during our sample period. The teens attend grade 5 in 2010, just after the school track has been chosen at the end of primary school (after grade 4) in most federal states. This cohort might be eligible through all four channels though the majority will not yet be individually eligible (as they are still under the age of 16). Finally, the third or *adolescent cohort* is on average 17 years of age and attend grade 9 at the beginning of our sample period, which marks the end of compulsory schooling in Germany. As these children are mostly born between 1995 and 1996, they are not eligible for birthright citizenship. Yet, they could be eligible under the transitional rule, through individual or parental eligibility.

Table 3.1 shows summary statistics of our sample of first-generation and second generation of immigrants in the NEPS. Around 85% of children in our sample were born in Germany. Those immigrant children who were born abroad came to Germany at the age of five on average.

3.3.2 Eligibility and Take-Up of Citizenship

To study the effect of citizenship on children's skill development and school performance, we will instrument actual German citizenship with the four access options to citizenship discussed in Section 3.2. In particular, we define separate binary indicators equal to one if a child with at least one foreign-born parent is eligible by birthright, under the transitional rule, or when they themselves or their parents are eligible for citizenship. The indicators are equal to zero if the child is not eligible under the specified access option. Eligibility

[13] While information on the main caretaker is available for most of the sample with less than 10% missing, the information on the second parent or partner is missing for 30% of our sample. We return to this issue in the robustness section below.

Table 3.1: Summary Statistics for the Sample of Immigrant Children

	Mean	Std. Dev.	Minimum	Maximum
Child Treatment:				
German Citizenship	0.81	0.39	0	1
Child Eligibility				
Birthright	0.40	0.49	0	1
Transitional Rule	0.34	0.48	0	1
Individual Eligibility	0.44	0.50	0	1
Parental Eligibility	0.96	0.19	0	1
Child Outcomes:				
Math	-0.16	0.96	-2.45	2.36
Vocabulary	-0.55	1.09	-4.69	2.62
Reading Comprehension	-0.11	0.97	-3.09	2.73
Science	0.12	0.97	-2	2.46
Computer	-0.08	0.99	-3	2.89
Grade in German	2.74	0.87	1	6
Grade in Math	2.80	1.05	1	6
Ever Repeatd a Grade	0.17	0.38	0	1
Academic Track Recommendation	0.44	0.50	0	1
Academic Track	0.42	0.49	0	1
Child Characteristics:				
Girl	0.50	0.50	0	1
Age	13.6	4.22	4.25	26
Birth Cohort	2000	4.70	1990	2008
Foreign-born	0.14	0.35	0	1
Age at Immigration if Foreign-born	4.9	3.61	0	16
Parental Characteristics:				
German Citizenship	0.59	0.49	0	1
Age	41.44	6.14	15	77
Years since Immigration	23.42	10.05	0	58
Birth Cohort	1971	6.71	1936	1996
Immigration Year	1990	10.13	1957	2012
Education: Low	0.33	0.47	0	1
Education: Medium	0.57	0.49	0	1
Education: High	0.10	0.30	0	1
Observations	23,868			

Notes: The sample is restricted to first- and second-generation immigrant children with at least one foreign-born parent. The eligibility variables are binary indicators equal to one if the child is eligible under one of the access options specified (see Institutional Background section in the main text). Parental education is coded as follows: low (no high school or vocational degree), medium (high school or vocational degree) and high (university or college degree).

Source: National Educational Panel Study, Starting Cohorts 2-4, 2010-2017

for any of the four access options depends on observable characteristics. Eligibility for birthright citizenship and the transitional rule depend on the child's country of birth (in Germany or abroad) and birth cohort as well as parental years since migration. Individual eligibility is defined based on the child's birth cohort and years since migration (or parental years since migration if the child is born in Germany). Finally, eligibility through the parents depends on parental years since migration and the child's birth cohort (as only children under age 18 can be included in the citizenship application).

Eligibility for citizenship will not be perfectly correlated with observed citizenship for at least two reasons: additional eligibility criteria and take-up.[14] All access options to citizenship depend on residency requirements of the parents or child. While fulfilling the residency requirement is an important eligibility criteria, it is not the only one. As discussed in Section 3.2, the absence of a severe criminal background or economic self-sufficiency, among others, are additional conditions to be fulfilled. Hence, some immigrants might satisfy the residency requirement but fail because of one of the other criteria. Even if they fulfill all eligibility criteria, immigrants might decide not to apply for German citizenship because of unobserved costs that exceed the benefits of obtaining host country citizenship. One potential source of such costs is that immigrants feel strongly attached to the source country, which might increase the perceived costs of naturalizing, especially if one needs to renounce the citizenship of the source country (Monscheuer 2019). Other factors might be social aspirations or motivations to apply for citizenship. All of these factors will be filtered out with our instrumental variables strategy.

Appendix Table 3.8 shows the share of immigrant children in our sample who are eligible under one of the four access options and their take-up of citizenship.[15] Among immigrant children born in Germany, the highest take-up rate (93%) occurs for those 47% in our sample who are eligible under birthright citizenship. Almost all (99%) children born in Germany to foreign-born parents have been eligible through their parents. 85% of those have naturalized. Foreign-born children in turn are most likely to be eligible through their parents (79%) but their take-up rates are with 64% much lower than for immigrant children born in Germany. Note that these take-up rates are substantially higher than for first-generation immigrants with more than ten years of residency in Germany, of which around 35-40% naturalized (OECD 2011).

[14] In addition, there can also be measurement error in the survey data. If it is classical measurement error, our estimates will be biased toward zero making it more difficult to detect causal effects of citizenship on skill development.

[15] Note that a child eligible for citizenship might not have naturalized under that specific category as some children are eligible through multiple channels. Take-up rates are similar if we restrict children to be eligible under one access option only (see notes to Table 3.8 for the actual take-up rates).

3.4 Econometric Framework

To analyze how host country citizenship affects educational outcomes and to explore the heterogeneity in treatment effects along observable and unobservable gains, we use the marginal treatment effects (MTE) framework developed by Björklund and Moffitt (1987), Heckman (1997) and Heckman and Vytlacil (1999, 2007b).

3.4.1 Setup

Let Y_{0i} be a potential outcome like test scores for child i in the non-treated and Y_{1i} the potential outcome for child i in the treated state. We model potential outcomes as linear in parameters:

$$Y_{ji} = X_i\beta_j + U_{ji} \qquad j = 0, 1 \tag{3.1}$$

where $E[U_j|X = x] = 0$, which is satisfied if we interpret equation (3.1) as a linear projection of Y_j onto X. The X denote child and parental characteristics that might influence educational outcomes even independently of citizenship.
The selection into treatment, i.e. whether a child has host country citizenship, is denoted by N_i. We follow the literature in modeling the treatment decision as a latent index model:

$$N_i^* = Z_i\beta_d - V_i \quad \text{with } N_i = 1 \text{ if } N_i^* \geq= 0 \text{ and } N_i = 0 \text{ otherwise} \tag{3.2}$$

where Z includes the same variables X as the outcome equation and in addition the instruments $\tilde{Z}$, which are excluded from the outcome equation. In our case, the excluded instruments $\tilde{Z}$ are binary indicators equal to one if a child is eligible for citizenship under a certain access option (birthright, transitional rule, individual or parental eligibility) and zero otherwise. The unobserved component of the selection equation V_i represents individual characteristics not observed by the econometrician that make a child less likely to obtain host country citizenship.
Without loss of generality, we can transform the selection equation (3.2), $Z_i\gamma_n - V_i \geq 0$, into $Z_i\gamma_n \geq V_i$ and $\Phi(Z_i\gamma_n) \geq \Phi(V_i)$ where Φ denotes the cdf of V (in our case, a standard normal). The term $\Phi(Z_i\gamma_n) \equiv P(Z)$ is the propensity score, while the second term $\Phi(V_i) \equiv U_N$ denotes a random variable distributed in the unit interval. Hence, a person naturalizes when the propensity score, which is in part determined by the available options to obtain citizenship, exceeds U_N, which represents the unobserved resistance to treatment. In what follows, we will think of U_N as the unobserved (monetary or psychic) costs of obtaining host country citizenship.
The gain from treatment for child i based on the potential outcomes in equation (3.1) is given as:

$Y_{i1} - Y_{i0} = X_i(\beta_1 - \beta_0) + (U_{i1} - U_{i0})$. The first term represents how gains vary with observable characteristics like gender, for example, in the treatment state, while the second term represents the unobserved gains from citizenship. Individuals might select into citizenship based on both observable and unobservable gains. As a result, the unobserved gains from citizenship ($U_{1i} - U_{0i}$) may be correlated with the unobserved costs of obtaining citizenship (U_N).

The marginal treatment effect (MTE) is then defined as:

$$MTE(X = x, U_N = u_N) = E[Y_1 - Y_0 | X = x, U_N = u_N] \tag{3.3}$$

The MTE represents the gain from treatment for an individual with observed characteristics $X = x$ and unobservable characteristics $U_N = u_N$. Hence, the MTE identifies the return to citizenship for an individual with $P(Z) = u_N$ who is just indifferent between naturalizing ($N = 1$) and not naturalizing ($N = 0$). Tracing the MTE at different U_N values reveals how the return to citizenship varies across different quantiles of the unobserved costs to obtain host country citizenship.

We make the following assumptions: (U_1, U_0, U_N) is independent of $(Z|X)$. In our context, this assumption implies that eligibility through one of the four access options $\tilde{Z}$ is as good as randomly assigned and hence, independent of the unobserved gains and the unobserved resistance to treatment conditional on the control variables X. We provide some evidence for this assumption in Section 3.4.3 below. Under the conditional independence assumption, we can trace the propensity score $P(Z)$ over the unit interval for a given X. In practice, however, we rarely have the data and variation in the instruments available, especially with a large set of control variables X (Carneiro et al. 2011) and binary instruments as in our case (Brinch et al. 2017). We therefore impose in addition the assumption that X and (U_1, U_0) are additively separable conditional on U_N (Brinch et al. 2017). [16] We can then trace the MTE over the unconditional support of the propensity score $P(Z)$, as opposed to the support of $P(Z)$ conditional on $X = x$. A consequence of this assumption is that the shape of the marginal treatment effect does not depend on X except for its intercept. Under additive separability, we can rewrite the MTE as

$$\text{MTE}(X = x, U_N = u_N) = E[Y_1 - Y_0 | X = x, U_N = u_N] = x(\beta_1 - \beta_0) + K(p) \tag{3.4}$$

where $K(p) = E[U_1 - U_0 | U_N = u_N]$. The first term represents the heterogeneous effect of treatment with respect to observable characteristics, while the second term characterizes the heterogeneity in unobserved gains to citizenship.

[16] These two assumptions are slightly weaker than the full independence assumption, i.e. (U_1, U_0, U_N) is jointly independent of $(\tilde{Z}, X)$ that is typically imposed in empirical work (Carneiro and Lee 2009; Carneiro et al. 2011; French and Song 2014; Maestas et al. 2013).

3.4.2 Empirical Specification and Estimation

The estimation proceeds in two steps. In the first stage, we estimate the selection equation, i.e. the empirical counterpart of equation (3.2). The dependent variable in the first stage is an indicator equal to one if a child has German citizenship and zero otherwise. The instruments $\tilde{Z}$ are the four eligibility indicators discussed in Section 3.3.2. Each indicator is equal to one if a child is eligible under the specific access option and zero otherwise.

As control variables X we include child and family characteristics that influence take-up and educational outcomes. As child characteristics we include gender, a linear and squared term in age (measured in years with monthly precision), an indicator whether the child was born abroad or in Germany and the child's age of arrival.[17] We also control for a full set of birth cohort fixed effects as the child's year of birth is an important determinant of eligibility. We further include parental age as this might affect both the financial resources available in the family as well as parenting styles like the time spent with the child.[18] A full set of parental years since migration fixed effects is included to absorb additional differences between immigrant parents that might affect the take-up or educational outcomes of the child.

In addition, we control for state fixed effects to adjust for state-level differences in educational policy as well as cohort- and wave-specific fixed effects. The latter ensure that we only compare outcomes between children belonging to the same cohort and wave. To adjust for differential propensities of take-up and school performance across source countries, we define ten regions of origin: the traditional EU-15 member states (e.g. Italy), immigrants from countries that recently joined the European Union (the EU-12, e.g. Poland), immigrants from Turkey, ex-Yugoslavia and the Former Soviet Union (except the Baltic states). We lump together other immigrants into broad regions of origin (Asia, Africa, the Middle East and North and South America).

We estimate the first stage as a probit model thus assuming that the cdf of V, Φ, is a standard normal. From the estimates of the citizenship decision, we predict the propensity scores for treated and non-treated individuals. Recall that our second-stage estimation of the MTE is identified non-parametrically only over the support of $P(Z)$. Therefore, we impose common support of $P(Z)$ for treated and non-treated individuals. In addition, we also trim 1% of the observations with the thinnest common support in the data. We show in Section 3.5.4 below that our results are robust to alternative trimming margins.

[17] Note that the child's age of arrival is equal to zero if the child was born in Germany. Age of arrival effects are therefore identified from children born abroad only.

[18] We do not include parental education, employment or number of siblings. While these are potentially important determinants of family resources and time spent with the child (Doepke et al. 2019), they are themselves influenced by citizenship eligibility of the parent (Gathmann and Keller 2018) or child (Avitabile et al. 2013, 2014).

In the second step, we estimate the outcome equation which is given by:

$$E[Y_i|X = x, U_N = u_N] = X_i\beta_0 + X_i(\beta_1 - \beta_0)p + K(p). \tag{3.5}$$

The MTE for individual i is then calculated as the derivative of equation (3.5) with respect to p.

We allow the observable gains to citizenship (the second term in equation (3.5)) to vary with child gender, whether the child is born abroad and parental age.[19] We also need to specify the functional form for $K(p)$, the unobserved gains to treatment (the third term in equation (3.5)). Our main results use a third-order polynomial in the propensity score, which restricts the MTE to be quadratic. In the robustness section, we demonstrate that more flexible, higher-order polynomials or semi-parametric methods to approximate $K(p)$ yield results that are similar to the baseline.

To estimate the second stage, we use local instrumental variables (Local Instrumental Variable (LIV)), which estimates equation (3.5) and then calculates its derivative with respect to p. Recently, Brinch et al. (2017) have propagated the separate estimation approach, which estimates the outcome equations in the treated and non-treated state separately and then calculates the MTE. This estimation approach can be particularly useful in the case of discrete instruments with limited variation. In our case, the four instruments of eligibility can take on twelve distinct combinations. Therefore, even using the LIV approach to estimate the MTE allows us to identify a flexible, high-order polynomial of the unobservable gains to treatment $K(P(Z))$.[20] We bootstrap all standard errors to account for the estimation error in the first stage.[21]

3.4.3 Exogeneity of Citizenship Eligibility

As discussed in Section 3.4.1, the identification of the MTE requires that the instruments $\tilde{Z}$ are jointly independent of unobservable gains in the outcome equation (U_{1i}, U_{0i}) and

[19] We include all other control variables as in the first stage but restrict its coefficients to be the same in the treated and untreated state.

[20] Note that the number is lower than $2^4 = 16$ because eligibility under birthright citizenship and the transitional rule are mutually exclusive. Suppose that the conditional expectations of the unobservables U_1 and U_0 are specified as parametric functions linear in L parameters. Even without any control variables the LIV approach can identify a MTE with $L \leq (M-1) = 11$ parameters, while the separate estimation approach can identify $L \leq M = 12$ parameters (see Proposition 1 in Brinch et al. 2017). Hence, LIV can still identify $K(P(Z))$ as a polynomial of order eleven in our case.

[21] While the instruments vary at a more aggregated level than the individual child, each instrument varies at a different level (e.g. individual eligibility depends on birth cohort and years since migration, while birthright eligibility depends on parental years since migration and whether the child is born in Germany). As a consequence, it is not obvious at which level to cluster or whether to cluster at all given that we include fixed effects for parental years since migration and the child's birth cohort (Abadie et al. 2017).

resistance to treatment in the selection equation (U_N) conditional on our control variables X. The four eligibility indicators, which serve as instrumental variables, only depend on observable, demographic characteristics like parental years since migration, the child's country of birth or birth cohort (see the discussion in Section 3.3.2). As such, the instruments are independent of any individual motivations and aspirations to succeed in the host country or attachment to the source country, which are likely determinants of the resistance to treatment (U_N).

The identifying assumption could still be violated if certain groups of eligible children benefit more from citizenship or have better preconditions to succeed. Immigrants from outside the EU who naturalize obtain easier access to certain jobs and professions, which has positive effects on earnings and educational investments (for a more detailed discussion see Gathmann and Keller 2018). For EU citizens, there are no restrictions as the single market guarantees the same access to jobs and professions as natives. If the composition of EU and non-EU immigrants coming to Germany changes over time (and hence, the composition of immigrants becoming eligible for citizenship), this could generate a spurious correlation between eligibility and educational outcomes. To rule this out, our baseline specification controls for the geographical composition of immigrants by including ten regions of origin fixed effects (including two categories for long-term and recent EU member states), as described in Section 3.4.2.

Furthermore, the returns to host country citizenship might also be higher or lower for foreign-born children than for children born and brought up in Germany. Similarly, the returns to citizenship might decline when parents have lived in the country for a long time and are well integrated in the host country. Both variables (being foreign-born and parental years since migration) also determine eligibility for at least one of the four access options, which might generate a spurious correlation between the instruments and unobservable gains to treatment. Finally, both age of arrival of first-generation immigrants and the child's birth cohort are correlated with eligibility: children arriving at a younger age have to wait longer for individual eligibility; the birth cohort in turn determines whether a child born in Germany is eligible for birthright citizenship, for instance. At the same time, it is well known that language acquisition is more difficult after a certain age, which would result in lower scores on tests measuring the skill in the host country language. And there might be cohort effects in skill development if schools or teachers improve over time or are better able to teach classes from diverse backgrounds. To control for these potential confounding influences, we include a full set of birth cohort and parental years since migration fixed effects as well as the child's age of arrival and whether the child is foreign-born as controls in our baseline specification.

Table 3.2 provides more direct evidence supporting the identifying assumption. Using

our instruments as dependent variables, we check whether observable characteristics are correlated with the instrument conditional on our set of control variables. As potential confounders we use parental employment and education, living arrangement and child gender. All of these are likely determinants of the returns to citizenship and resistance to treatment. The results in Table 3.2 show that out of the 24 coefficients, only two are statistically significant at the 5% level and a third one at the 10% level. The F-statistic reported at the bottom of the table shows that we can reject the joint significance of the regressors for all four instruments.

Another concern could be that selective return migration could bias our estimates. Specifically, if immigrants are less likely to return to their home country if they become eligible faster or have access to birthright citizenship for their children, for instance, our sample of eligible children differs from the sample of not (or not yet) eligible children. Depending on the type of return migration, this could bias our estimates upward or downward. We would obtain an upward bias, for instance, if non-eligible children with the worst educational performance are more likely to leave the host country. We return to this issue in the robustness section.

Table 3.2: Balancing Tests

	Eligibility by Birth (1)	Transitional Rule (2)	Individual Eligibility (3)	Eligibility through Parent (4)
Child is a Girl	0.003	0.003	-0.002	-0.000
	[0.003]	[0.005]	[0.004]	[0.000]
Partner living in Household	-0.001	0.010	-0.002	-0.001**
	[0.005]	[0.009]	[0.006]	[0.001]
Parental Employment	0.002	-0.010	0.005	0.000
	[0.004]	[0.006]	[0.004]	[0.001]
Mother Working Full-Time	0.006	-0.003	-0.002	0.000
	[0.004]	[0.007]	[0.005]	[0.001]
Parental Education: medium	-0.007	0.015**	-0.007*	-0.001
	[0.004]	[0.006]	[0.004]	[0.001]
Parental Education: high	-0.009	0.003	-0.005	0.000
	[0.006]	[0.009]	[0.007]	[0.001]
F-Test Joint Significance of Covariates	1.04	1.91	0.70	0.83
(p-value)	0.40	0.10	0.65	0.55

Notes: The table reports regression estimates where each column comes from a separate regression. The dependent variables shown in the top row are binary indicators equal to one if a child is eligible for citizenship by birth (column (1)), eligible under the transitional rule (column (2)), individually eligible (column (3)) or eligible through parents (column (4)) and zero otherwise. A parent is medium-skilled if she has a high school or vocational degree and she is high-skilled if she holds a college or university degree. The omitted category is low-skilled, i.e. without a high school or vocational degree. The p-value shown is for the F-test of joint significance of the covariates shown in the table. All specifications also include whether a child is born abroad, child age and age squared as well as its age of arrival (which is coded as zero if the child is born in Germany), birth cohort fixed effects, age of the parent and fixed effects for the number of years the parent has lived in Germany. Robust standard errors are reported in square brackets. * p<0.1, ** p<0.05 and *** p<0.01.

Source: National Educational Panel Study, Starting Cohorts 2-4, 2010-2017

3.5 Empirical Results

3.5.1 Selection into Citizenship

We start out with the estimates of the selection equation (3.2). Table 3.3 shows marginal effects from a probit model where the dependent variable is equal to one if the child holds German citizenship and zero otherwise. The first column shows the results for the whole sample, while columns (2) to (4) report estimates for the subsamples in which the language tests were conducted. The estimates show that the four access options have a strong effect on the likelihood of citizenship. Birthright citizenship, which is available for children born in Germany only, raises the probability of having a German passport by 14.6 percentage points in the full sample. Individual eligibility, which is available when immigrant children turn sixteen is also an important channel to obtain citizenship, in particular for children born abroad. Compared to children born in Germany, foreign-born children are about 10 percentage points more likely to naturalize under individual eligibility.[22]

Recall that children eligible under the transitional rule are, by definition, eligible through their parents as well. Hence, the negative coefficient on the transitional rule thus implies that children are less likely to naturalize under the transitional rule than children only eligible through their parents. This effect is in line with aggregate statistics that show a very low take-up under the transitional rule, most likely because many immigrant parents did not know about this transitional policy. Eligibility through parents plays a minor role for citizenship in our context as almost all children (96%) in our samples are eligible through their parents. The F-statistic at the bottom of Table 3.3 confirms that the four access options have a strong effect on selecting into treatment (citizenship) across all samples. Foreign-born children are overall less likely to obtain citizenship unless they become individually eligible. We also find strong negative effects of age and age at immigration on take-up. In contrast, there are no gender differences in the likelihood of citizenship acquisition.

As we can identify returns to citizenship at quantiles of U_N within the support of the distribution of $P(Z)$ only, Figure 3.3 plots the distribution of estimated propensity scores for treated and untreated children. The common support generated by both instruments and observable characteristics covers the range between 0.47 and 0.97.[23] Hence, we cannot

[22] While the overall pattern is similar across samples, the coefficients are typically larger in magnitude in the three subsamples. The main reason is that eligibility options differ across birth cohorts and grades.

[23] Figure 3.3 shows common support for the whole sample. In Figure 3.6, we also show the propensity scores for treated and untreated children in the subsamples with valid test scores for reading and vocabulary (in Panel A and B) or information on German grades (in Panel C). The common support

say much for immigrants who are most likely to take up citizenship unless we are willing to extrapolate based on functional form assumptions like normality, for instance. Given aggregate take-up rates of between 40-80% among all immigrants (OECD 2011), the common support generated by our instruments covers an important, (policy-)relevant range of take-up rates.[24]

Table 3.3: Selection Equation

	Sample (Full) (1)	Sample (Reading) (2)	Sample (Vocabulary) (3)	Sample (German Grade) (4)
Birthright Citizenship	0.146***	0.459**	0.793***	0.455***
	[0.017]	[0.185]	[0.211]	[0.169]
Transitional Rule	-0.104***	-0.654***	-0.856***	-0.533***
	[0.012]	[0.108]	[0.152]	[0.104]
Individual Eligibility	-0.006	-0.339*	-0.825**	0.208
	[0.014]	[0.176]	[0.382]	[0.143]
Individual Eligibility*Foreign-born	0.100***	0.533***	0.936***	0.548***
	[0.010]	[0.172]	[0.213]	[0.147]
Eligibility through Parents	-0.051**	-0.168	4.622	-0.042
	[0.026]	[0.682]	[105.789]	[0.544]
Foreign-born	-0.171***	-0.597***	-0.653***	-0.612***
	[0.025]	[0.178]	[0.189]	[0.152]
Child is a Girl	0.003	0.018	0.097*	0.071
	[0.005]	[0.052]	[0.059]	[0.047]
Age	-0.065***	-0.534*	0.159	-1.282***
	[0.020]	[0.275]	[0.363]	[0.312]
Age Squared	0.001	0.002	-0.012	0.038***
	[0.001]	[0.011]	[0.021]	[0.012]
Age at Immigration	-0.009***	-0.113***	-0.103***	-0.114***
	[0.002]	[0.028]	[0.035]	[0.025]
Parental Age	-0.001	-0.006	-0.011**	0.003
	[0.000]	[0.005]	[0.006]	[0.004]
Partial R2 Instruments	0.019	0.017	0.031	0.013
F-Test Joint Significance of Instruments	64.0	11.5	275.4	10.00
(p value)	0.000	0.000	0.000	0.000
Observations	23,065	3,920	3,465	4,609

Notes: The table reports average marginal effects of a probit selection model where the dependent variable indicates whether the child has German citizenship or not. Column (1) uses the full sample across all survey cohorts and waves, while columns (2)-(4) use the sample with valid tests in reading comprehension (column (2)), vocabulary (column (3)) and German grade (column (4)). The instruments are four indicators equal to one if a child is eligible for citizenship under the respective access option and zero otherwise. These are birthright citizenship (child born in Germany after December 31, 1999 with at least one foreign-born parent who satisfies eight years of legal residency at birth); the transitional rule (child born in Germany between January 1, 1990 and December 31, 1999 with at least one foreign-born parent who satisfies eight years of legal residency and applies for the child in 2000); individual eligibility (child is at least 16 and has lived legally in the country for at least eight years); and eligibility through parents (child is under 18 and has one foreign-born parent who satisfies the residency requirement (eight years for immigrants arriving in 1992 or later, eight years for teen immigrants arriving under the age of 15 before 1992; or up to fifteen years for adults arriving at the age of 15 or older prior to 1992)). We also interact individual eligibility with an indicator whether the child is born abroad to allow for differential take-up rates. Age at immigration of the child is equal to zero for children born in Germany. In addition to the variables shown, all specifications further include birth cohort fixed effects, parental years since migration fixed effects and survey cohort interacted with wave-specific fixed effects. Robust standard errors are reported in square brackets. *** p<0.01, ** p<0.05 and * p<0.1.

Source: National Educational Panel Study, Starting Cohorts 2-4, 2010-2017

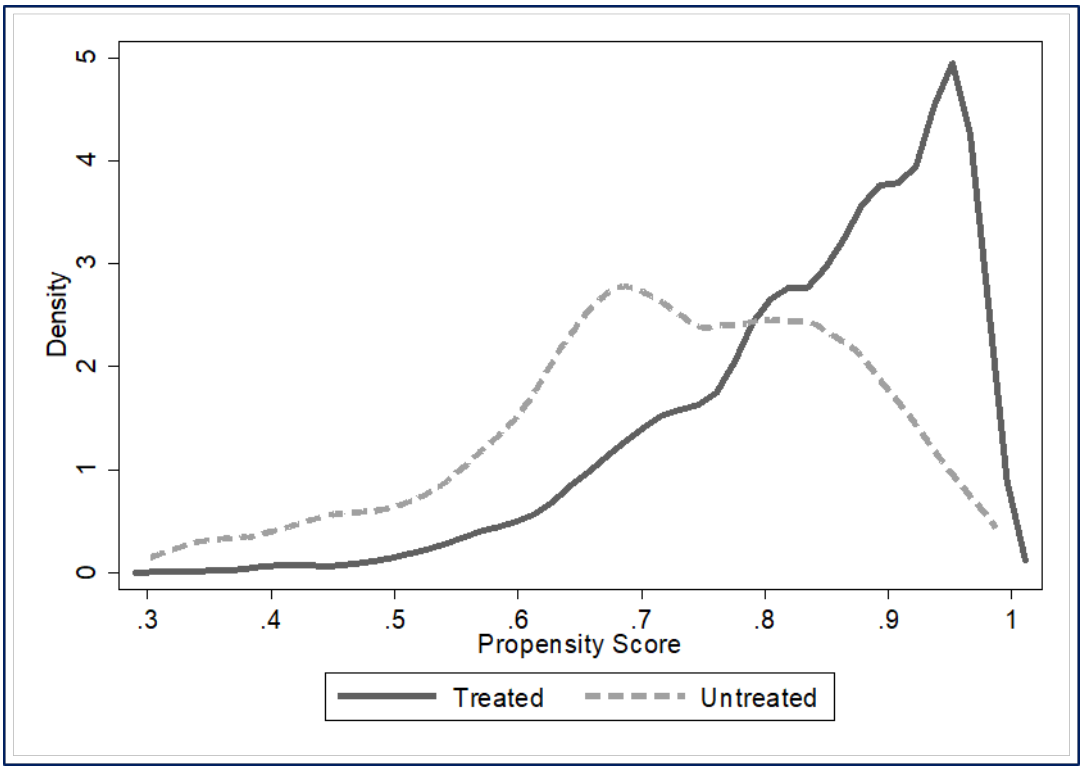
3.5.2 Language Skills

We now turn to the question whether and how citizenship affects language skills as measured by standardized test scores in vocabulary and reading and school grades in German.

for reading comprehension (in Panel A) and German grade (in Panel C) is [0.47; 0.97], for vocabulary (in Panel B) the common support covers the slightly broader range [0.42; 0.98].

[24] We are not the only study estimating heterogeneous treatment effects with limited common support in the propensity score: Maestas et al. (2013) obtain coverage between 0.57 and 0.85 and French and Song (2014) cover the range between 0.45 and 0.85 for disability receipt, while Felfe and Lalive (2018) have coverage between 0 and 0.5 of the propensity to attend early childhood education.

Figure 3.3: Common Support for Treated and Untreated Individuals



The results, based on estimating equation (3.5), in Table 3.4 show that citizenship allows some immigrant children to improve their skills relative to their peers. Girls who do not obtain German citizenship have a vocabulary that is 0.61 of a standard deviation below their male peers. Obtaining citizenship allows girls to close the gap in language skills compared to immigrant boys. They also improve their German language grade by 0.3 grades.

Foreign-born children, in turn, have better language skills in the untreated citizenship than children born in Germany. Obtaining citizenship allows immigrant children who are born in Germany to catch up with foreign-born children. Similarly, immigrant children with older parents have a better vocabulary in German than children with younger parents, possibly because parents spend more time to improve their child's command of the German language. Obtaining citizenship neutralizes the advantage of parental age and allows children with younger parents to catch up to their peers.

Overall, the evidence from Tables 3.3 and 3.4 suggests that foreign-born children and children with older parents are less likely to obtain citizenship and also have lower returns to citizenship in terms of language skills. Immigrant girls who have higher gains in terms of language skills are somewhat more likely to select into treatment than boys. These results suggest a pattern of positive selection into citizenship in terms of observed characteristics. This finding is in line with estimates for adult immigrants where more

educated immigrants are more likely to naturalize (see e.g. Chiswick and Miller 2008; Gathmann and Keller 2018).

Figure 3.4 plots the MTE curve for reading comprehension (Panel A), vocabulary (Panel B) and German grade (Panel C) at the mean values of observable characteristics (X) in the sample. The figures relate the gains from treatment ($U_1 - U_0$) on the y-axis to the unobserved costs of citizenship (U_N) on the x-axis. For language skills, the MTE curve is downward sloping: individuals who face higher costs (or higher resistance) to citizenship have lower gains than individuals with a low unobservable resistance. Individuals with low resistance to citizenship have positive gains, while individuals with high resistance have zero returns to citizenship.

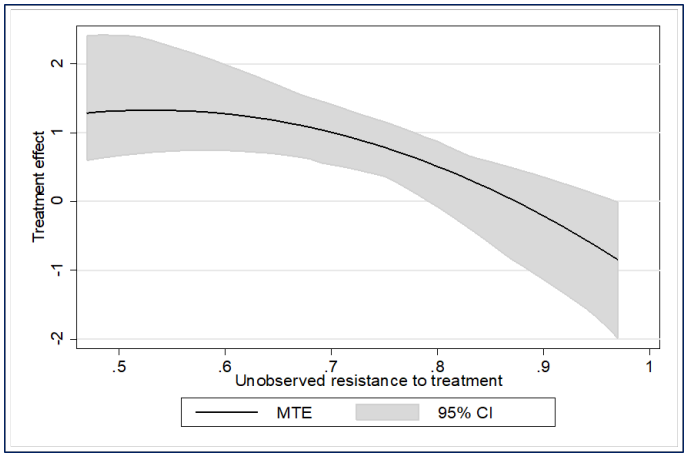
The pattern for reading comprehension and vocabulary points to positive selection into citizenship in terms of unobservables. Panels A and B in Figure 3.4 suggest that children with an unobserved resistance smaller than $U_N < 0.8$ and $U_N < 0.75$, respectively have statistically significant positive returns to citizenship, while children above these resistances have zero returns. The slope of the MTE curve and hence, heterogeneity in returns in terms of unobservables is statistically significant at the 2% level for reading comprehension and at the 1% level for vocabulary (see the p-value of the test for essential heterogeneity reported in column (1) and (2) of Table 3.4). The estimates for school grades in German are also consistent with positive selection. Recall that school grades vary from one to six with one being the best and six the poorest grade. Panel C of Figure 3.4 shows that immigrant children with the lowest resistance to treatment have positive treatment effects, i.e. experience an improvement in grades. Unobservable gains are zero for children with $U_N \geq 0.6$. This threshold when unobservable gains are zero is lower for German grades than for standardized test scores.

Overall then, the positive selection into citizenship with respect to language skills implies that those with the highest returns are most likely to take-up citizenship. Returns are lower and approach zero for immigrant children at the margin of take-up, i.e. with unobservable characteristics that make them more hesitant to take up citizenship. We will return to the question whether a further liberalization of citizenship law generates additional benefits in the next section below.

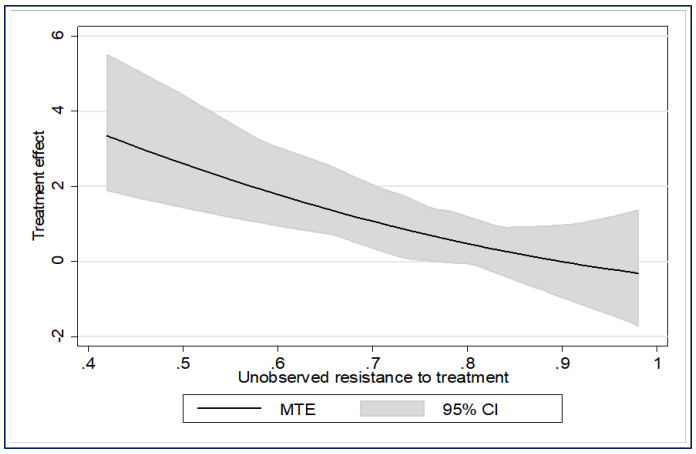
3.5.3 Other Estimates and Outcomes

So far, we have shown that immigrant children with the highest propensity to become German citizens have the highest returns in terms of improved German language skills. Understanding the language of instruction and knowing how to express oneself in class is important to follow and understand other subjects taught. We next investigate whether citizenship has positive effects on the performance in other subjects and school more

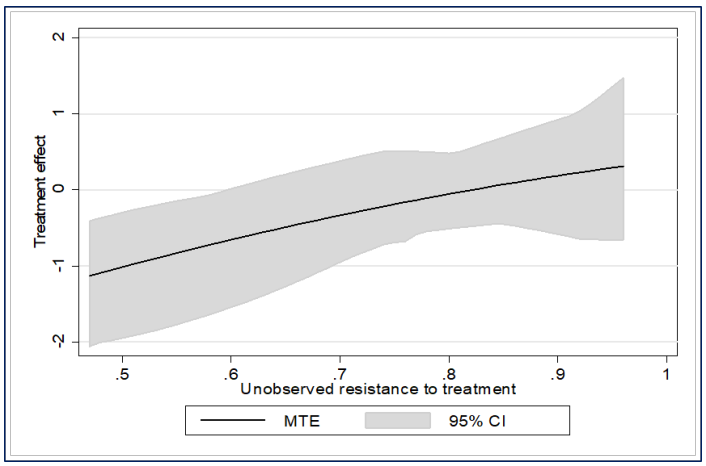
Figure 3.4: Marginal Treatment Effects for Language Skills



Panel A



Panel B



Panel C

Table 3.4: MTE Estimates for Test Scores

	Reading (1)	Vocabulary (2)	German Grade (3)
Child is a Girl	-0.218	-0.617	-0.018
	[0.16]	[0.21]***	[0.15]
Child Foreign-born	0.627	1.378	-0.724
	[0.36]*	[0.47]***	[0.42]*
Parental Age	0.011	0.053	-0.003
	[0.01]	[0.01]***	[0.01]
Propensity Score	11.930	-14.517	5.809
	[6.04]**	[10.69]	[6.20]
Propensity Score Squared	-11.248	5.706	-2.003
	[4.56]**	[8.16]	[5.10]
Propensity Score x Girl	0.366	0.527	-0.305
	[0.19]*	[0.25]**	[0.18]*
Propensity Score x Foreignborn	-0.792	-1.373	0.968
	[0.53]	[0.59]**	[0.54]*
Propensity Score x Parental Age	0.004	-0.048	-0.010
	[0.02]	[0.02]***	[0.02]
Test Observable Heterogeneity (p-value)	0.017	0.001	0.048
Test for Essential Heterogeneity (p-value)	0.019	0.004	0.060
Observations	3,920	3,465	4,609

Notes: The table reports estimates from the outcome equation where the dependent variables are standardized test scores in reading comprehension (column (1)), vocabulary in German (column (2)) and German grade (column (3)). Coefficients not interacted with the propensity score measure the effect in the untreated state, while coefficients interacted with the propensity score represent the difference in effect between treatment (host country citizenship) and no treatment. In addition to the variables shown, all specifications also include child age and age squared, the child's age at immigration, birth cohort fixed effects and parental years since migration fixed effects (all not interacted with the propensity score). Bootstrapped standard errors are reported in square brackets. *** p<0.01, ** p<0.05 and * p<0.1.

Source: National Educational Panel Study, Starting Cohorts 2-4, 2010-2017

broadly. Our data provide us with standardized test scores for math, natural sciences and computer skills. In addition, we know the grade in math, whether a child ever repeated a grade in school, whether the child attends the academic high school track and whether the primary school teacher recommended the child for the academic track or not.[25]

Table 3.5 shows the results of estimating our baseline model in equations (3.2) and (3.5) where the dependent variables are cognitive tests and other measures of school performance. Given that we only have limited support, we do not report global treatment parameters (like ATE or Average Treatment Effect on the Treated (ATT)) as these would depend on strong functional form assumptions like joint normality.[26] Instead we report the LATE as well as marginal policy-relevant treatment effects (Marginal Policy Relevant Treatment Effects (MPRTE)). The latter are obtained by perturbing the propensity score by a factor α and assessing its effect on the returns to citizenship for the marginal child between treatment and no treatment (for details see Carneiro et al. 2011). The basic assumption is that a marginal change to the propensity score (with α approaching zero) only shifts $P(Z)$, the propensity to obtain citizenship, but does not affect potential outcomes or the unobservables determining selection (U_1, U_0 and U_N). Hence, knowing how the distribution of the propensity score P^α changes is sufficient to calculate a policy-relevant treatment effect. As we cannot identify the full support of the propensity score, we estimate MPRTEs, which identify how the returns to citizenship for immigrant children change at the margin of take-up. Our first perturbation adds a fixed amount α to each child's propensity score (hence $P_\alpha = P + \alpha$). The resulting MPRTE1 is then the MTE aggregated over individuals (with observables X and unobservables U_N) who would switch into treatment in response to the perturbation. The second perturbation is to multiply each child's propensity score with $(1 + \alpha)$, so $P_\alpha = P(1 + \alpha)$ and calculating the MPRTE2. In practice, the two methods yield very similar results.

As a baseline for comparison, columns (1) to (3) of Table 3.5 show the estimates for our main measures of language skills. The LATE, which is a weighted average of the MTE over the relevant distribution of X and U_N is positive but does not reach statistical significance at conventional levels. The MPRTE are positive for test scores and statistically significant

[25] In most states, parents decide on the school track for their child, which may or may not coincide with the recommendation of the primary school teacher. In some states (Baden-Württemberg until 2012, Bavaria, Brandenburg, Saxony, Saxony-Anhalt and Thuringia), the recommendation by the primary school teacher prescribed the track choice. In all six states, parents can circumvent (or overrule) the prescribed track choice by requesting a slot or trial period in the academic track.

[26] Alternatively, we could estimate average treatment or average treatment on the treated effects by rescaling the weights to sum to one over the common support (see Carneiro et al. 2011). For reading comprehension, the ATE, ATT and Average Treatment Effect on the Untreated (ATUT) (with bootstrapped standard errors in brackets) are 0.682 [0.26]***, 0.946 [0.27]*** and -0.018 [0.42] respectively. For vocabulary, we get as ATE 1.22 [0.36]***, as ATT 1.571 [0.41]*** and as ATUT 0.004 [0.46]. These effects might differ, however, from the effect of an individual outside the common support.

in three out of four cases suggesting that expanding take-up of citizenship would improve the language skills of the marginal child.

Columns (4) to (6) of Table 3.5 show estimates for test scores in math, sciences and ICT skills. For all three outcome variables, the LATE estimates are positive, but only reach statistical significance for math skills at the 10% level. The MPRTE are also positive for math skills indicating that expanding take-up would improve math skills of the child at the margin. Interestingly, the test of essential heterogeneity at the bottom of Table 3.5 shows that we cannot reject the null hypothesis of no unobservable gains of citizenship for math and science skills. The test statistic for ICT skills indicates in turn that there is some heterogeneity but reaches statistical significance only at the 15% level. The corresponding MTE curves in Panels A (math), B (sciences) and C (ICT skills) of Figure 3.7 illustrate that pattern. For math skills, Panel A suggests that the effects of citizenship are positive and sizable across the distribution of unobservable resistance to treatment. As the MTE curve is basically flat, there is no sizable heterogeneity in unobservable gains that varies with U_N. For sciences skills, the MTE graph shows some curvature but the variation is not enough to reject the null of no essential heterogeneity.

These findings indicate that either there is no heterogeneity in unobservables for math and science skills or agents (i.e. parents and children) are not aware of it and hence, do not act on these gains when deciding whether to apply for citizenship. Econometrically, the absence of essential heterogeneity implies that, under additive separability, equation (3.4) becomes $E(Y_1 - Y_0|X = x, U_N = u_N) = E(Y_1 - Y_0|X = x)$, so $Y_1 - Y_0$ is mean independent of U_N given $X = x$. In that case, all the treatment parameters are the same, so MTE = LATE.[27]

Finally, we turn to the question whether immigrant children with host country citizenship perform better in school. Columns (7) to (10) of Table 3.5 show modest LATE effects for grades in math, whether a child attends the academic track and whether the primary school teacher recommended the academic track or not. Similarly, the MPRTE are close to zero and never statistically significant. The p-value at the bottom further indicates that there is no heterogeneity in returns to citizenship with respect to resistance to treatment. The only exception is grade retention (shown in column (8) of Table 3.5): immigrant children with German citizenship are much less likely to repeat a grade in school. The p-value at the bottom also suggests that the null hypothesis of no essential heterogeneity can be rejected at the 5% level.

[27] In principle, we would have MTE = ATE = ATT = LATE. Yet, given our limited common support, we cannot rule out the possibility that there is essential heterogeneity outside the common support for those most or least likely to take-up citizenship.

Table 3.5: Other Educational Outcomes

	Language Skills			Other Test Scores			School Performance			
	Reading Comprehension	Vocabulary	German Grade	Math	Natural Sciences	Computer Skills	Math Grade	Grade Retention	Academic Track	Recommendation
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
LATE	0.278	0.347	0.049	0.492	0.190	0.361	0.159	-0.227	0.075	-0.160
	[0.34]	[0.27]	[0.29]	[0.25]*	[0.34]	[0.30]	[0.34]	[0.09]**	[0.12]	[0.15]
MPRTE 1	0.024	0.113	-0.010	0.031	0.003	0.008	0.001	-0.010	0.002	-0.003
	[0.01]*	[0.03]***	[0.01]	[0.02]**	[0.01]	[0.01]	[0.01]	[0.01]	[0.01]	[0.01]
MPRTE 2	0.011	0.061	-0.006	0.028	-0.001	-0.011	0.002	-0.011	-0.001	-0.009
	[0.01]	[0.03]*	[0.01]	[0.02]	[0.01]	[0.02]	[0.01]	[0.01]	[0.02]	[0.01]
p-value Ess. Het.	0.019	0.004	0.060	0.894	0.338	0.134	0.709	0.032	0.839	0.184
Observations	3,920	3,465	4,609	5,839	2,854	3,088	4,603	3,047	2,233	1,858

Notes: The table reports estimates from the outcome equation where the dependent variables are language skills (standardized test scores in reading comprehension in column (1) and vocabulary in column (2)), standardized test scores (math in column (4), natural sciences in column (5) and computer skills in column (6)) as well as measures of school performance (grade in German and math in columns (3) and (7), whether a child has ever repeated a grade in column (8), whether a child attends the academic track (high school) in column (9), and whether the primary school teacher recommended the academic track or not in column (10)). The LATE is calculated by aggregating the MTE over the range of observed and unobserved characteristics of compliers, i.e. those individuals that are induced to change from no treatment to treatment when the instruments are turned on. We also report two marginal policy-relevant treatment effects: the MPRTE1 is the treatment effect if we perturb the propensity score of each individual by Pa=P+a where a=0. MPRTE2 is the treatment effect if we perturb the propensity score of each individual by Pa=(1+a)*P where a=0. MPRTE1 and MPRTE2 then identify how citizenship affects educational outcomes for those induced to switch into treatment under the perturbed propensity score. In addition, we report the p-value of the null hypothesis that there is no essential (unobserved) heterogeneity. The specification is the same as for the baseline reported in Tables 3.3 and 3.4. Bootstrapped standard errors are reported in square brackets. *** p<0.01, ** p<0.05 and * p<0.1.

Source: National Educational Panel Study, Starting Cohorts 2-4, 2010-2017

3.5.4 Robustness

The pattern of positive selection in language skill gains is robust to a number of alternative specifications. Our baseline estimation uses a second-order polynomial to approximate $K(P(Z))$, which implies a linear MTE function. Hence, unobserved gains can either monotonically increase or decrease over the quantiles of U_N but not both. To allow for more flexible MTE curves over the distribution of U_N, we estimate models with a third- or fourth-order polynomial where the dependent variables are standardized test scores in reading comprehension. The results are shown in Panel A of Figure 3.5. Allowing the MTE curve to be quadratic (using a third-order polynomial) is very similar to the baseline: both linear and quadratic MTE curves are downward sloping in line with positive selection on unobserved gains. Allowing for an even more flexible fourth-order polynomial still yields a downward sloping MTE curve, which becomes upward sloping for those least likely to take-up citizenship ($U_N > 0.9$). Using a semiparametric approach, we again find a very similar downward sloping shape of the MTE curve. Overall then, the positive selection in unobservable gains is robust to alternative functional form assumptions.

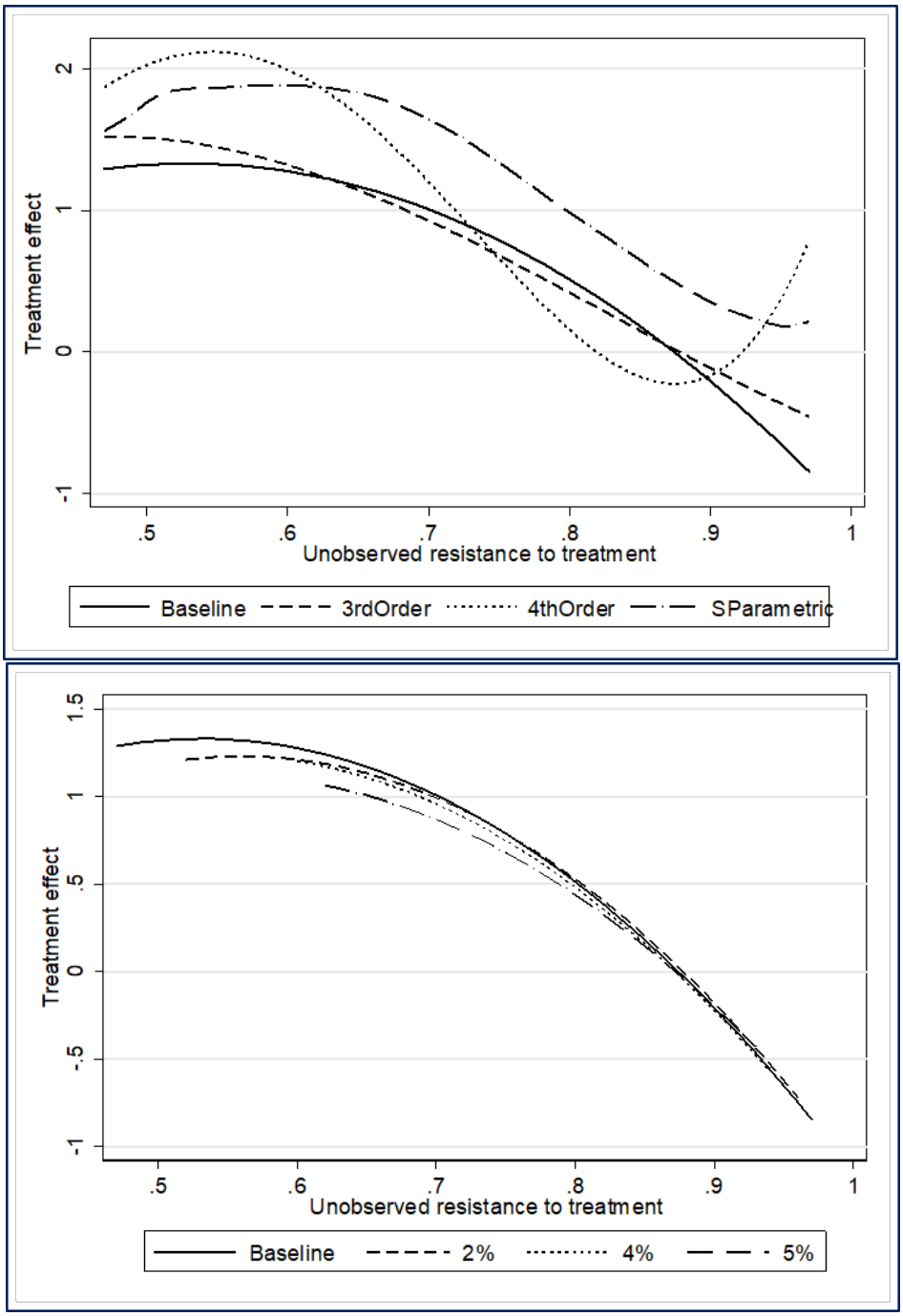
In our baseline, we also trim 1% observations with the thinnest common support. In Panel B of Figure 3.5, we report estimates using alternative trimming margins of 2%, 4% and 5%. Additional trimming mostly cuts observations at low levels of resistance (with $U_N < 0.6$). The resulting downward sloping MTE curves, however, are very similar across alternative trimming margins. That suggests that our results are not driven by observations with little common support.

Table 3.6 reports a number of additional robustness checks for our main outcomes reading comprehension in Panel A and vocabulary in Panel B. The first column shows the baseline estimates for comparison. We again report LATE and MPRTE effects. The MPRTE are calculated based on the same two perturbations of the propensity score we discussed in the last section. We first check the sensitivity of our results to alternative specifications of the outcome equations. In column (2), we implement an alternative semi-parametric estimator based on a double residual regression approach (Robinson 1988).[28] Column (3) uses the separate approach, which estimates the outcome equation in the treated and untreated state separately, rather than local instrumental variables (LIV). In both cases, we get even stronger positive effects on language skills in the baseline.

The next set of robustness checks uses alternative specifications of the selection equation. In column (4), we specify the treatment equation as a logit rather than a probit. In column (5), we only use our four discrete instruments of eligibility without interacting

[28] In the first step, we obtain residuals from regressions of Y, X and Xp on polynomial in p. In the second step, we estimate β_0 and $\beta_1 - \beta_0$ by regressing the residualized outcome on the residualized regressors. We then use a local polynomial regression of the residualized outcomes on p to obtain an estimate of $K(p)$. MTE is then calculated by taking the derivative of $K(p)$.

Figure 3.5: Specification Checks



individual eligibility with an indicator for being foreign-born. Including an interaction effect seems a natural specification as foreign-born children do not have access to birthright citizenship. We would therefore expect that foreign-born children are more likely to take advantage of individual or parental eligibility. We drop the interaction effect in column (5) to rule out that the naturalization effect in the first stage is primarily identified through the interaction effects rather than the instruments themselves. The results show that functional form and interaction between instruments and controls have little effect on our estimates, which are very similar to the baseline in column (1).

As children's cognitive skills typically improve with age, one might worry that our linear and squared term are not enough to pick up age-related differences in language capacities. In column (6) we, therefore, add a full set of birth month dummies to our baseline. The LATE effects for reading comprehension are somewhat stronger although the effects on the marginal child are very similar to in the baseline. For vocabulary we find typically even stronger positive effects for the set of compliers as well as the marginal child. One might also worry that the observable or unobservable characteristics of subsequent immigrant cohorts change over time and are correlated with both treatment and potential outcomes. In addition to our years since migration fixed effects, we therefore include in column (7) cohort of arrival fixed effects in 5-year bands (to avoid multicollinearity between years since migration, year of arrival and survey year). The estimates indicate again strongly positive returns to citizenship for the set of compliers as well as the marginal child.[29]

Finally, one might worry that parents adjust their fertility behavior in response to the citizenship reform. In particular, parents may have delayed conception or birth to ensure that their child is eligible for birthright citizenship. If parents delay conception, we should see a spike in birth in January of 2000 and a (corresponding) decline of births in December of 1999. Vital statistics, however, exhibit no discontinuity in the number of immigrant children born in the time period before or after January 1, 2000 (see Appendix Figure A1 in Felfe et al. (2019)). Parents scheduled to deliver around the cutoff date could have tried to delay the birth of their child until January 1, 2000. To rule this out, we drop from our sample children born in December of 1999 and January of 2000 in column (8) of Table 3.6. This additional restriction reduces the sample by only 50 children. The results are again somewhat stronger for reading comprehension and very similar for vocabulary compared in the baseline in column (1).[30]

[29] To check whether outliers in terms of parental age, year of arrival or child age have any impact on our results, we also drop parents born before 1970 (N=1,514), parents arriving in Germany prior to 1975 (N=453) or children older than 19 (N=11) in additional checks. In all cases the results obtained are similar to the baseline.

[30] Parents might also adjust their long-run fertility by reducing the number of children (see Avitabile et al. 2014: for evidence on birthright citizenship), having children at a later age (see Gathmann et al. 2019: for evidence on parental eligibility) and investing more in each child in line with the

Across the many different specifications in Table 3.6, the p-values reported below the estimates in each panel indicate that we cannot reject the null hypothesis of no essential heterogeneity in only three out of the sixteen specifications at the 15% level.

quantity-quality tradeoff (Becker and Lewis 1973; Becker and Tomes 1976). Such an adjustment in parental behavior in response to the opportunities of host country citizenship are fully in line with the mechanisms discussed in the introduction.

Table 3.6: Robustness Checks

	Baseline MTE (1)	Semi-Parametric (2)	Separate Estimation (3)	Logit 1st Stage (4)	No Interaction 1st Stage (5)	Birth Month Dummies (6)	Parental YSM (7)	Drop 12/99 and 01/00 (8)
Panel A: Reading								
LATE	0.278	0.797	0.463	0.248	0.474	0.851	0.995	0.908
	[0.34]	[0.15]***	[0.26]*	[0.27]	[0.31]	[0.26]***	[0.28]***	[0.23]***
MPRTE 1	0.024	0.040	0.024	0.013	0.026	0.023	0.048	0.021
	[0.01]*	[0.01]***	[0.01]**	[0.02]	[0.01]**	[0.01]*	[0.01]***	[0.01]**
MPRTE 2	0.011	0.029	0.019	0.010	0.016	0.012	0.029	0.011
	[0.01]	[0.01]***	[0.01]*	[0.01]	[0.01]	[0.01]	[0.01]**	[0.01]
Essential Heterogeneity	0.02	0.00	0.38	0.01	0.11	0.10	0.21	0.02
Observations	3,920	3,920	3,920	3,920	3,917	3,918	3,919	3,873
Panel B: Vocabulary								
LATE	0.347	0.550	0.461	0.425	0.436	0.561	0.852	0.319
	[0.27]	[0.23]**	[0.26]*	[0.29]	[0.37]	[0.29]*	[0.26]***	[0.31]
MPRTE 1	0.113	0.109	0.074	0.117	0.055	0.151	0.187	0.105
	[0.03]***	[0.02]***	[0.02]***	[0.07]*	[0.02]**	[0.04]***	[0.04]***	[0.03]***
MPRTE 2	0.061	0.059	0.040	0.067	0.030	0.087	0.109	0.058
	[0.03]*	[0.02]***	[0.02]*	[0.03]*	[0.02]	[0.03]***	[0.03]***	[0.03]**
Essential Heterogeneity	0.00	0.00	0.09	0.00	0.01	0.01	0.01	0.02
Observations	3,465	3,465	3,465	3,465	3,437	3,465	3,465	3,459

Notes: The dependent variable in Panel A is the standardized test score in reading comprehension; in Panel B, the dependent variable is the standardized test score in German vocabulary. The table shows the LATE and two marginal policy-relevant treatment effects (MPRTE1 and MPRTE2). The LATE is calculated by aggregating the MTE over the range of observed and unobserved characteristics of compliers, i.e. those individuals that are induced to change from no treatment to treatment when the instruments are turned on. We also report two marginal policy-relevant treatment effects: the MPRTE1 is the treatment effect if we perturb the propensity score of each individual by Pa=P+a where a=0. MPRTE2 is the treatment effect if we perturb the propensity score of each individual by Pa=(1+a)*P where a=0. MPRTE1 and MPRTE2 then identify how citizenship affects educational outcomes for those induced to switch into treatment under the perturbed propensity score. In addition, we report the p-value of the null hypothesis that there is no essential (unobserved) heterogeneity. Column (1) reports the effects for the baseline MTE using the same specification than in Tables 3.3 and 3.4. Column (2) estimates the MTE semiparametrically using double residual regression. Column (3) estimates the outcomes in the treated and non-treated state separately. Column (4) uses a logit to estimate the first-stage selection equation. Column (5) only includes the four instruments without interacting it with foreignborn. Column (6) adds a set of birth month dummies to adjust for small age differences between children attending the same grade. Column (7) adds fixed effects of the parent's arrival cohort (in 5-year bands). Column (8) drops children born in December of 1999 and January 2000 to reduce concern about the selective delay of birhs (donut estimate). Bootstrapped standard errors are reported in square brackets. *** p<0.01, ** p<0.05 and * p<0.1.

Source: National Educational Panel Study, Starting Cohorts 2-4, 2010-2017

3.6 Policy Simulations

The higher returns for children with the lowest resistance to treatment raises the question whether a reform of immigration policy would carry any additional net benefits. To explore this possibility, we calculate policy-relevant treatment effects (PRTE), which characterize the effect of a change from the existing policy to an alternative policy regime (Heckman and Vytlacil 2001, 2005; Carneiro et al. 2011).

The basic idea is to consider alternative policies that shift $P(Z)$, the propensity to obtain citizenship, but do not affect potential test scores or the unobservables determining selection (U_1, U_0 and U_N). Hence, knowing how the policy changes the distribution of the propensity score P^α is sufficient to calculate the policy-relevant treatment effect of the reform considered. As we cannot identify the full support of the propensity score, we cannot estimate policy-relevant treatment effects that allow for arbitrary variation in the take-up of citizenship. Instead, we estimate marginal versions of the PRTE, which identify how small changes in take-up or eligibility might affect the returns to citizenship for immigrant children.

For each alternative policy, we generate a counterfactual distribution of propensity scores. Using the estimated coefficients from the second stage, we aggregate the MTE over those individuals who switch treatment states in the alternative policy regime.[31] We explore the consequences of three alternative policy regimes. Our first counterfactual experiment is to increase the likelihood of take-up. We perform the policy simulation in two ways: the first counterfactual exercise adds 10 percentage points to the propensity score of each child up to a maximum of one (hence, we set $P^\alpha = P + \alpha$ where $\alpha = 0.1$). The second counterfactual raises the propensity score of each child by 10% (by setting $P^\alpha = P(1+\alpha)$ where again $\alpha = 0.1$). The results are contained in row (1) of Table 3.7. Encouraging take-up among immigrant children has few additional gains in terms of language skills irrespective of how we perturb the propensity score.[32] With positive selection in gains and high take-up (83% in our case), those induced to obtain citizenship in the counterfactual scenario, have low or no returns to citizenship in this scenario.

Our second counterfactual scenario is a more restrictive citizenship policy that reduces take-up. One way to achieve this to raise the actual or implicit costs of obtaining citizenship (increase in application fees or waiting time) or by tightening eligibility criteria (e.g. by requiring more than six years of basic education, for instance). Similar to the first scenario, we implement this in two ways: by *subtracting* 10 or 20 percentage points

[31] See Table 1 in Carneiro et al. (2011) for the corresponding weights to calculate PRTE from MTE for alternative policy changes.

[32] Note that the propensity score does not increase by 0.1 to 0.93 in the first counterfactual because we set the maximum to one.

from the propensity score or multiplying the propensity score by 0.9 and 0.8 respectively. The results show that restricting take-up has positive effects in terms of language skills. A small reduction in take-up does not reach statistical significance (see row (2a) in Table 3.7). In contrast, a stricter access to citizenship for immigrant children, which would reduce take-up by up to 20 percentage points, has positive gains. Hence, inducing some children not to obtain host country citizenship, would actually be beneficial in terms of their language skills.

Our third counterfactual scenario is one where we implement a policy that grants birthright citizenship to all second-generation immigrants irrespective of their parents' years of residency in Germany. This counterfactual scenario is thus close to the one followed in traditional immigration countries like the United States or Canada, for instance. To implement this counterfactual exercise, we set the eligibility indicator for birthright citizenship to one for all immigrant children born in Germany. We then recalculate the distribution of the propensity score and the policy-relecant treatment effect. Row (3) in Table 3.7 shows that such a policy reform would imply a modest increase in take-up from 82% to 87% but imply sizable gains in terms of improved language skills.

Table 3.7: Simulations of Alternative Citizenship Policies

	Propensity Score			Policy-Relevant Treatment Effects		
	Sample	Counterfactual 1	Counterfactual 2	Pa = P + a	Pa = (1+a)*P	Z[k]a = Z[k] +a
(A) Increase Citizenship Take-up (by 10pp or 10%)	0.820	0.899	0.883	0.178 [0.392]	0.157 [0.368]	
(B1) Restrict Access to Citizenship (by 10pp or 10%)	0.820	0.720	0.737	0.415 [0.319]	0.337 [0.279]	
(B2) Restrict Access to Citizenship (by 20pp or 20%)	0.820	0.621	0.656	0.631 [0.300]**	0.572 [0.268]**	
(C) Give Birthright Citizenship to all Second-Generation Immigrants	0.820	0.866				0.528 [0.236]**

Notes: The table presents estimates of policy-relevant treatment effects on reading comprehension (measured by standardized test scores) from three different policy experiments (described in the first column). The first policy experiment raises take-up of citizenship (row (A)); the second experiment restricts take-up (row (B1) and (B2)). The third experiment grants automatic citizenship by birth to all immigrants born in Germany (row (C)). To implement the first and second policy experiment, we calculate the PRTE for two alternative perturbations of the propensity score: in the first case, we add X percentage points to each propensity score (MPRTE1). The second perturbation raises the propensity score of each individual by X% (MPRTE2). For (A1). Column (1) shows the take-up rate in the sample, while columns (2) and (3) report the counterfactual take-up rates in the alternative policy regime. Columns (4) and (5) then report the policy-relevant treatment effect (PRTE) of the alternative policy, which is calculated by aggregating the Marginal Treatment Effect (MTE) for individuals induced to switch treatment in the alternative policy regime. Column (6) shows the PRTE when we set eligibility for birthright citizenship, one of our instruments, to one for all second-generation immigrants (MPRTE3). Bootstrapped standard errors are shown in square brackets. *** p<0.01, ** p<0.05 and * p<0.1.

Source: National Educational Panel Study, Starting Cohorts 2-4, 2010-2017

3.7 Conclusion

Granting access to citizenship is one of the most fundamental integration policies a host country can offer. In this paper, we estimate marginal returns (MTE's) to citizenship for children of immigrants. We study two important policy reforms in Germany, a country that has substantially liberalized its citizenship policy in recent decades.

The four different access options to citizenship introduced by the policy reforms in 1991 and 2000, birthright citizenship, an associated transitional rule, eligibility through parents and individual eligibility, enable us to explore naturalization decisions in much detail. We find offering birthright citizenship to be the most important access channel, while there are also interesting effects of individual eligibility. Here, children react differently by place of birth. Becoming eligible for naturalization in their own right encourages foreign-born children much more successfully to naturalize than German-born children.

The offer to naturalize signals to immigrants that they fully belong to the host country society, therefore, granting citizenship should encourage further investments in important host country specific skills, such as local language. Exploring the effect of citizenship on German language abilities in more detail, we indeed uncover promising effects. Granting citizenship enables three important groups to overcome their weaker performances in language tests or grades − immigrant girls, German-born children and children of younger parents. Immigrant girls catch up to immigrant boys, who otherwise display more elaborate vocabulary skills than girls. Also, German-born children are able to close the gap to foreign-born kids in terms of reading abilities, vocabulary test results and German grades. Usually, immigrant children with older parents display better German vocabulary than children with younger parents, possibly because parents spent more time to improve their child's command of the German language. Obtaining citizenship, too, allows children of younger parents to catch up and close the gap to children of older parents in terms of vocabulary. Interestingly, we find evidence for positive selection into citizenship for children of immigrants. This is in contrast to previous studies analyzing the effects of the same German reforms for grown-up immigrants. Gathmann and Keller (2018), for instance, find intermediate selection into citizenship for men and even negative selection for women.

Understanding the language of instruction and knowing how to express oneself in class is important to follow and understand all school subjects. Therefore, one might expect improving language skills would lead to improvements in other skills or school subjects as well. Interestingly, we only observe an improvement of math skills and a decrease in the probability of grade retention, whereas most other skills like natural science or computer abilities as well as further measures of school performance seem to be unaffected.

The fact that we observe positive selection into treatment, i.e. higher returns for children with lowest resistance to treatment, raises the question whether a reform of immigration policy would carry any additional net benefits. We explore this in three ways. Firstly, we conduct a policy simulation that encourages citizenship take-up and find that this has few additional gains. Due to the positive selection in gains and high-take up, those induced to obtain citizenship have low or no returns to citizenship. Secondly, we simulate a

policy that reduces take-up. The results show that restricting take-up has positive effects in terms of language skills. In our third counterfactual scenario we implement a policy that grants birthright citizenship to all second-generation immigrants irrespective of their parent's year of residency in Germany. This counterfactual scenario is thus close to the one followed in traditional immigration countries. We find that such a policy reform in Germany would imply a modest increase in take-up but imply sizeable gains in terms of improved language skills.

Overall, this paper provides encouraging evidence on returns to citizenship for children, yet, results on objective tests, grades and policy simulations point out that attention to detail is key when deciding on the specific policy design.

3.A Appendix

3.A.1 The National Educational Panel Study

The National Educational Panel Study (NEPS) provides a comprehensive set of standardized competence tests (for a comprehensive overview see Weinert et al. 2011). We focus on general competences such as reading, vocabulary, math, natural sciences and ICT literacy, which are administered to all students. We have up to three measurements of the language and math skills and up to two measurements for the evolution of a child's science and ICT skills. The tests are mostly pen-and-paper, cover age-adjusted topics and levels of difficulties, and take about 30 minutes to complete. Tests contain between 21 to 89 items and each item is given the same weight. To make the tests comparable across age groups, we standardize test scores to values between 0 and 100 in each starting cohort and consider test outcomes in terms of standard deviations. Usually, two to three of these competence tests are conducted in each wave. Assessments are conducted on two consecutive days for younger children to reduce test fatigue.

Language skills: NEPS tests German language skills along several dimensions. We focus in our analysis on tests for reading comprehension and vocabulary as these are available for all cohorts in multiple waves (for a more detailed discussion see Berendes et al. 2013). *Reading comprehension* tests consist of handling different text formats and comprehension. The first part of the NEPS reading competence tests participant's ability to understand and interpret texts encountered in everyday situations: texts providing information or instructions, discursive or literary texts, or advertisements. The test approach is very similar to other large-scale education tests like PISA.[33] In addition, NEPS tests the cognitive requirements to handle texts like finding information in the text or drawing text-related conclusions and assessments. Test items are either provided as multiple choice, true-false answers or by matching the appropriate item like a headline to a text passage. Tests take approximately 30 minutes and are conducted in a pen and paper-based in-classroom situation for most age groups. Very young children undergo an adapted testing procedure. They are tested in a one-to-one situation with an adult interviewer and tests are situation oriented. More details on test content and test development for reading comprehension can be found in Weinert et al. (2011) and Gehrer et al. (2013). *Vocabulary* tests are based on the Peabody Picture Vocabulary Test (PPVT, Dunn and Dunn (1981)) and similar to those in other large-scale panel studies (e.g. British Cohort Study or the European Child Care and Education Study). More detailed information on tests for receptive vocabulary, i.e. listening comprehension at the word level, can be found in Hecker et al. (2015) and Berendes et al. (2013).

Math skills: The math tests are closely related to OECD's Programme for International Student Assessment (PISA) and Germany's Mathematics Education Standards (GMES). The goal of the latter is to harmonize educational standards across the sixteen federal states, which have been developed by the National Council of Teachers of Mathematics (Weinert et al. 2011). NEPS tests both the mathematical knowledge and skills to meet real-life challenges (as covered by the international PISA test) and the age-specific

[33] Yet, in contrast to the PISA test, NEPS does not include discontinuous text formats containing diagrams, graphs or tables because these require additional skills not closely related to reading comprehension.

skills that students should master in a certain grade (according to the school curriculum). Test items cover areas like quantity, change and relationships, space and shape, data and chance, but also logical communication, argumentation and modeling using representational forms, mathematical problem solving, technical abilities and skills. Generally, the math tests take 30 minutes and are conducted in a pen and paper-based multiple-choice format in a class-room situation. Children in pre-K (kindergarten or first grade) undergo an adapted testing procedure. They are tested in a one-to-one situation with an adult interviewer and tests are situation oriented. A typical task in a kindergarten math test would be as follows: the child is asked for the number of stones an interviewer has placed in a bowl. The interviewer then adds a certain number of stones (always less than 10) and asks the child again for the correct number of stones. More detailed information including examples of test items can be found in Neumann et al. (2013), Weinert et al. (2011), and Ehmke et al. (2009).

Science skills: The NEPS scientific literacy test uses concepts from the PISA test program, the American Association for the Advancement of Science and the German educational standards. The test covers knowledge of science (like knowledge of basic scientific concepts and facts) and knowledge about science (like the understanding of scientific processes). For each cohort, the final test takes 30 minutes and consists of 23 to 26 items. For younger age groups (in kindergarten and first grade) tests are picture based and conducted in a one-to-one testing situation with an adult interviewer. The interviewer leads the children through a story (e.g. a kindergarten party where the child is asked to help solving scientific problems) and asks them to choose the correct answer from a picture. Tests are pen-and-paper multiple-choice tests and are conducted in a class-room situation. In rare cases, computer- or web-based tests are conducted instead (Fuß et al. 2016). Details on test procedures and test development can be found in Weinert et al. (2011) and Hahn et al. (2013).

ICT literacy: The Test of Technological and Information Literacy (TILT) tests knowledge and skills needed for a problem-oriented use of modern ICT. The underlying test concepts comes from the PISA test framework and the International ICT Literacy Panel (Digital Transformation: A Framework for ICT Literacy, International ICT Literacy Panel (2002)). The test contains items on technological literacy (define, access, manage, create), communication literacy (integrate, evaluate, communicate) and five different software applications (word processing, spreadsheets, presentation software, e-mail, search engines). Tests are given as either pen-and-paper or computer- and web-based. A typical test item, for instance, consists of a spreadsheet table with daily ticket sales. Participants are then provided with a multiple choice answer for the formula to calculate weekly ticket sales (for further details see Senkbeil et al. 2013; Weinert et al. 2011).

3.A.2 School Performance

We also use information on grades, track choice and grade retention to asses children's progress in school. These measures are interesting because they are informative about how teachers and educational staff assess immigrant children with or without host country citizenship. Yet, these dimensions are also more likely to vary across schools and across states.

School Grades: Each year, students are asked about the mark in their last school report.

Grades range from 1 (very good) to 6 (insufficient) where a grade of 4 or better is required to pass the subject.
Grade Retention: Grade retention is quite common in Germany affecting about one in six students over their school career. We use an indicator variable whether a student has ever repeated a grade or not. As this information is not collected annually, we use the most recent wave available for each student up to grade 5 for the child cohort (which we first observe at age 4 in pre-K) and up to grade 10 for the teen and adolescent cohort (which we observe since grade 5 and 9 respectively). We generally exclude all observations beyond grade 10 because only students in the academic track to obtain a high school degree (A-levels) remain in schools of general education, whereas the rest enters vocational training.
Track Recommendation and Choice: We code a variable equal to one if a student attends the academic track, which leads up to a high school degree (A-levels). Based on teacher recommendations, parents decide whether, after finishing primary school, their child goes to Hauptschule (finishing after grade 9), Realschule (finishing after grade 10) or High School (finishing with a high school degree after grade 12/13). In most states, parents can go divert from the recommendation of teachers (except in six states where teacher recommendations are binding). We also code a binary indicator whether the teacher recommended academic track (leading up to a high school degree) or not. Both variables on track choice are available in grade 5 for the child and teen cohorts and in grade 9 for the adolescent cohort as we only observe them from grade 9 onward.

3.A.3 Additional Tables

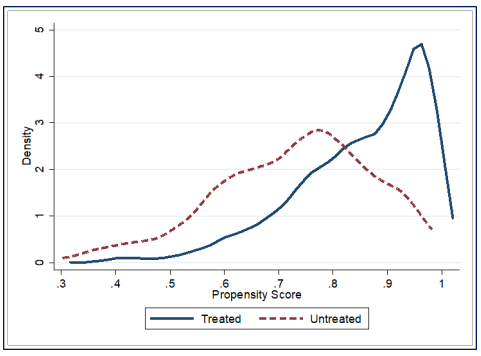
Table 3.8: Citizenship Eligibility and Take-Up in NEPS Sample

	Children born Abroad (First-generation immigrants)	Children born in Germany (Second-generation immigrants)
Citizenship by Birth	0%	47% (93%)
Transitional Rule	0%	40% (75%)
Individual Eligibility	55% (77%)	42% (69%)
Eligibility through Parent	79% (64%)	99% (85%)

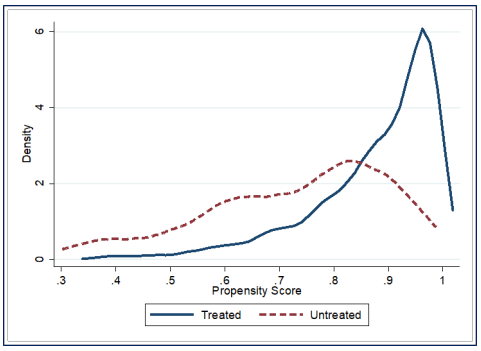
Notes: The table shows the percentage of children in the NEPS sample eligible under the four access options separately for children born in Germany and abroad. The numbers in parentheses show the percentage of children who are eligible under the specified access option with German citizenship („take-up"). Conditioning on the set of children not eligible under any other access option, the take-up rates are for foreign-born children: 75% rather than 77% under individual eligibility and 55% rather than 64% for eligibility through parents; and for children born in Germany 66% rather than 69% under individual eligibility and 84% rather than 85% for eligibility through parents.

3.A.4 Additional Figures

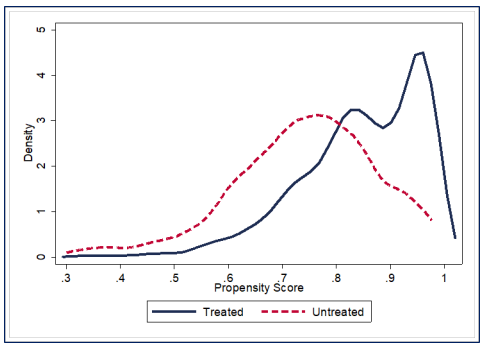
Figure 3.6: Common Support for Subsamples with Test Scores



Panel A

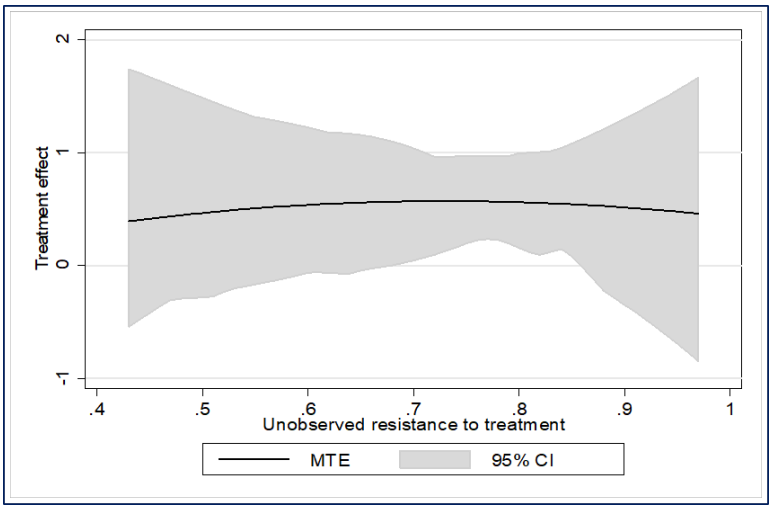


Panel B

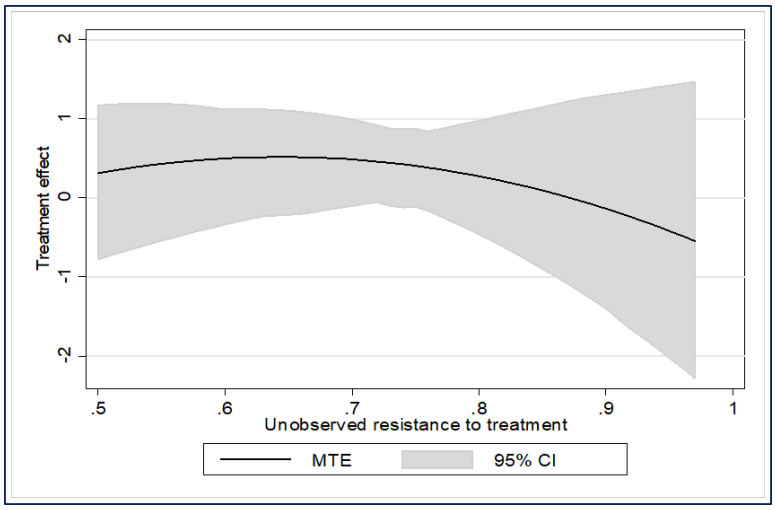


Panel C

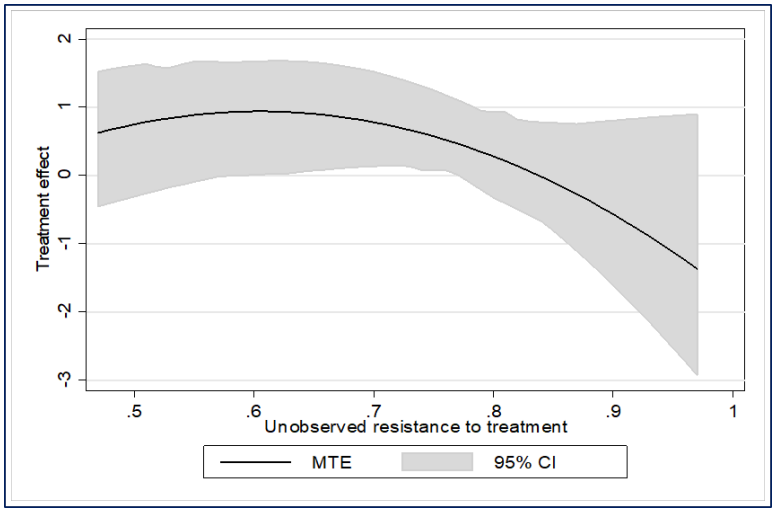
Figure 3.7: Marginal Treatment Effects for Math and Science Skills



Panel A



Panel B



Panel C

Bibliography

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017). When should you adjust standard errors for clustering? NBER Working Paper No. 24003, National Bureau of Economic Research.

Acemoglu, D. and A. Finkelstein (2008). Input and technology choices in regulated industries: evidence from the health care sector. *Journal of Political Economy 116*(5), 837–880.

Aiken, L. H., S. P. Clarke, D. M. Sloane, J. Sochalski, and J. H. Silber (2002). Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *Journal of the American Medical Association 288*(16), 1987–1993.

Aiken, L. H., C. SP, C. RB, S. DM, and S. JH (2003). Educational levels of hospital nurses and surgical patient mortality. *Journal of the American Medical Association 290*(12), 1617–1623.

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

Angrist, J. D., S. H. Chen, and J. Song (2011). Long-term consequences of vietnam-era conscription: new estimates using social security data. *The American Economic Review 101*(3), 334–338.

Athey, S. and G. W. Imbens (2017). The state of applied econometrics: causality and policy evaluation. *Journal of Economic Perspectives 31*(2), 3–32.

Attanasio, O., H. Low, and V. Sanchez-Marcos (2008). Explaining changes in female labour supply in a life-cycle model. *American Economic Review 98*(4), 1517–52.

Avitabile, C., I. Clots-Figueras, and P. Masella (2013). The effect of birthright citizenship on parental integration outcomes. *Journal of Law and Economics 56*(3), 777–810.

Avitabile, C., I. Clots-Figueras, and P. Masella (2014). Citizenship, fertility, and parental investments. *American Economic Journal: Applied Economics 6*(4), 35–65.

Baker, M. (2011). Innis lecture: Universal early childhood interventions: what is the evidence base? *Canadian Journal of Economics 44*(4), 1069–1105.

Baker, M., J. Gruber, and K. Milligan (2008). Universal childcare, maternal labor supply, and family well-being. *Journal of Political Economy 116*(4), 709–745.

Bartel, A. P., N. D. Beaulieu, C. S. Phibbs, and P. W. Stone (2014). Human capital and productivity in a team environment: evidence from the healthcare sector. *American Economic Journal: Applied Economics 6*(2), 231–59.

Bauer, T. K., S. Bender, A. R. Paloyo, and C. M. Schmidt (2012). Evaluating the labor-market effects of compulsory military service. *European Economic Review 56*(4), 814

– 829.

Bauer, T. K., S. Bender, A. R. Paloyo, and C. M. Schmidt (2014). Do guns displace books? The impact of compulsory military service on educational attainment. *Economics Letters 124*(3), 513 – 515.

Bauernschuster, S. and M. Schlotter (2015). Public child care and mothers' labor supply - Evidence from two quasi-experiments. *Journal of Public Economics 123*, 1–16.

Becker, G. S. and H. G. Lewis (1973). On the interaction between the quantity and quality of children. *Journal of Political Economy 81*(2, Part 2), 279–288.

Becker, G. S. and N. Tomes (1976). Child endowments and the quantity and quality of children. *Journal of Political Economy 84*(4, Part 2), 143–162.

Beher, K., P. Cloos, M. Galuske, R. Liebig, and T. Rauschenbach (2002). *Zivildienst und Arbeitsmarkt*, Volume 222 of *Schriftenreihe des Bundesministeriums für Familie, Senioren, Frauen und Jugend*. Stuttgart: Verlag W. Kohlhammer.

Berendes, K., S. Weinert, S. Zimmermann, and C. Artelt (2013). Assessing language indicators across the lifespan within the German National Educational Panel Study (NEPS). *Journal for Educational Research Online 5*(2), 15–49.

Berger, E., O. Groh-Samberg, and K. Spiess (2008). Die öffentlich geförderte Bildungs- und Betreuungsinfrastruktur in Deutschland: Eine ökonomische Analyse regionalspezifische Unterschiede. Innocenti Working Paper IWP-2008-03, UNICEF Office of Research - Innocenti.

Berlinski, S. and S. Galiani (2007). The effect of a large expansion of pre-primary school facilities on pre-school attendance and maternal employment. *Labour Economics 14*, 665–680.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics 119*(1), 249–275.

Björklund, A. and R. Moffitt (1987). The estimation of wage gains and welfare gains in self-selection models. *Review of Economics and Statistics 69*(1), 42–49.

Black, S. E., P. Devereux, and K. G. Salvanes (2011). Too young to leave the nest? The effects of school starting age. *Review of Economics and Statistics 93:*(2), 455–67.

Black, S. E., P. J. Devereux, K. V. Løken, and K. G. Salvanes (2014). Care or cash? The effect of child care subsidies on student performance. *Review of Economics and Statistics 96*(5), 824–837.

Blanden, J., E. DelBono, S. McNally, and B. Rabe (2016). Universal pre-school education: the case of public funding with private provision. *Economic Journal 126*, 682–723.

Blossfeld, H.-P., H.-G. Roßbach, and J. E. von Maurice (2011). Education as a lifelong process - the German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft Special Issue 14, Suppl 2*.

Bonin, H., A. Fichtl, H. Rainer, K. Spiess, and K. Wrohlich (2013). Zentrale Resultate der Gesamtevaluation familienbezogener Leistungen. DIW Wochenbericht No. 40, DIW.

Bratsberg, B., J. F. Ragan, and Z. M. Nasir (2002). The effect of naturalization on wage growth: a panel study of young male immigrants. *Journal of Labor Economics 20*(3), 568–597.

Breyer, F., P. Zweifel, and M. Kifmann (2013). *Gesundheitsökonomik*. Heidelberg:

Springer.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017). Beyond LATE with a discrete instrument. *Journal of Political Economy 125*, 985–1039.

Cameron, C. A. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources 50*(2), 317–372.

Card, D. and A. R. Cardoso (2012). Can compulsory military service raise civilian wages? Evidence from the peacetime draft in Portugal. *American Economic Journal: Applied Economics 4*(4), 57–93.

Card, D. and T. Lemieux (2001). Going to college to avoid the draft: the unintended legacy of the vietnam war. *The American Economic Review 91*(2), 97–102.

Carneiro, P. and R. Ginja (2014). Long term impacts of compensatory pre-school on health and behavior: evidence from Head Start. *American Economic Journal: Economic Policy 6*(4), 135–73.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011). Estimating marginal returns to education. *American Economic Review 101*(6), 2754–2781.

Carneiro, P. and S. Lee (2009). Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. *Journal of Econometrics 149*, 191–208.

Carneiro, P., K. Loeken, and K. G. Salvanes (2015). A flying start? Maternity leave benefits and long-run outcomes of children. *Journal of Political Economy 123*(2), 365–412.

Carneiro, P., M. Lokshin, and N. Umapathi (2017). Average and marginal returns to upper secondary schooling in Indonesia. *Journal of Applied Econometrics 32*(1), 16–36.

Cascio, E. U. (2009). Maternal labor supply and the introduction of kindergartens into American public schools. *Journal of Human Resources 44*(1), 140–170.

Cascio, E. U. (2017). Does universal preschool hit the target? Program access and pre-school impacts. NBER Working Paper No. 23215, National Bureau of Economic Research.

Cawley, J., D. C. Grabowski, and R. A. Hirth (2006). Factor substitution in nursing homes. *Journal of Health Economics 25*(2), 234–247.

Chiswick, B. and P. W. Miller (2008). Citizenship in the US: the roles of immigrant characteristics and country of origin. *Research in Labor Economics 29*, 91–130.

Cook, A., M. Gaynor, M. Stephens Jr, and L. Taylor (2012). The effect of a hospital nurse staffing mandate on patient health outcomes: evidence from california's minimum staffing regulation. *Journal of Health Economics 31*(2), 340–348.

Cookson, R., M. Laudicella, and P. L. Donni (2013). Does hospital competition harm equity? Evidence from the English National Health Service. *Journal of Health Economics 32*(2), 410–422.

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (2018). Who benefits from universal childcare? Estimating marginal returns to early childcare attendance. *Journal of Political Economy 126*(6), 2356–2409.

Cunha, F. and J. Heckman (2007). The technology of skill formation. *American Economic Review 97*(2), 31–47.

Cunningham, S. A., K. Mitchell, K. V. Narayan, and S. Yusuf (2008). Doctors' strikes and mortality: a review. *Social Science & Medicine 67*(11), 1784–1788.

Currie, J. and D. Thomas (1995). Does Head Start make a difference? *American Economic Review 85*(3), 341–364.

Datta Gupta, N. and M. Simonsen (2010). Non-cognitive child outcomes and universal high quality child care. *Journal of Public Economics 94*(1-2), 30–43.

De Los Reyes, A. and A. Kazdin (2005). Informant discrepancies in the assessment of childhood psychopathology: a critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin 131*(4), 483–509.

Deaton, A. and N. Cartwright (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine 210*, 2–21.

Destatis (2016). *Gesundheit - Ausgaben*. Fachserie 12 Reihe 7.1.1. Wiesbaden: Destatis.

Destatis (2017). *Grunddaten der Krankenhäuser*. Fachserie 12: Gesundheitswesen, Reihe 6.1.1. Stuttgart: Metzler-Poeschel.

Doepke, M., G. Sorrenti, and F. Zilibotti (2019). The economics of parenting. *Annual Review of Economics 11*, 55–84.

Doyle, J. J., S. M. Ewer, and T. H. Wagner (2010). Returns to physician human capital: evidence from patients randomized to physician teams. *Journal of Health Economics 29*(6), 866–882.

Dunn, L. M. and L. M. Dunn (1981). *Peabody Picture Vocabulary Test-revised*. Circle Pines, MN: American Guidance Service/Pearson Education.

Dustmann, C. (2008). Return migration, investment in children, and intergenerational mobility. Comparing sons of foreign- and native-born fathers. *Journal of Human Resources 43*(2), 299–324.

Dustmann, C. and A. Glitz (2011). Migration and education. In E. A. Hanushek, S. Machin, and L. Woessmann (Eds.), *Handbook of the Economics of Education*, Volume 3, pp. 327–439. Amsterdam: Elsevier.

Dustmann, C. and U. Schönberg (2012). The effect of expansions in maternity leave coverage on children's long-term outcomes. *American Economic Journal: Applied Economics 4*(3), 190–224.

Ehmke, T., C. Duchhardt, H. Geiser, M. Grüßing, A. Heinze, and F. Marschick (2009). Kompetenzentwicklung über die Lebensspanne - Erhebung von mathematischer Kompetenz im Nationalen Bildungspanel. In A. Heinze and M. Grüssing (Eds.), *Mathematiklernen vom Kindergarten bis zum Studium. Kontinuität und Kohärenz als Herausforderung für den Mathematikunterricht*, pp. 313–327. Münster: Waxmann Verlag GmbH.

Entorf, H. and N. Minoiu (2005). What a difference immigration policy makes: a comparison of PISA scores in Europe and traditional countries of immigration. *German Economic Review 6*(3), 355–376.

Evans, W. N. and B. Kim (2006). Patient outcomes when hospitals experience a surge in admissions. *Journal of Health Economics 25*(2), 365–388.

Felfe, C., M. G. Kocher, H. Rainer, J. Saurer, and T. Siedler (2018). More opportunity, more cooperation? The behavioral effects of birthright citizenship on immigrant youth.

Technical report, CESifo Working Paper No. 6991.

Felfe, C. and R. Lalive (2018). Does early child care help or hinder child development? *Journal of Public Economics 159*, 33–53.

Felfe, C., H. Rainer, and J. Saurer (2019). Why birthright citizenship matters for immigrant children: impacts on parental educational choices. *Journal of Labor Economics forthcoming*.

Fisher, A. C. (1969). The cost of the draft and the cost of ending the draft. *The American Economic Review 59*(3), 239–254.

Fitzpatrick, M. D. (2010). Pre-schoolers enrolled and mothers at work? The effects of universal pre-kindergarten. *Journal of Labor Economics 28*(1), 51–85.

Fleischhauer, J. (2007). *Wehrpflichtarmee und Wehrgerechtigkeit: Die Verfassungsmäßigkeit der allgemeinen Wehrpflicht im Blickwinkel sicherheitspolitischer, gesellschaftlicher und demographischer Veränderungen*. Hamburg: Dr. Kovač.

Fort, M., A. Ichino, and G. Zanella (2019). The cognitive cost of daycare 0-2 for children in advantaged families. *Journal of Political Economy forthcoming*.

French, E. and J. Song (2014). The effect of disability insurance receipt on labor supply. *American Economic Journal: Economic Policy 6*(2), 291–337.

Friedrich, B. U. and M. B. Hackmann (2017). The returns to nursing: evidence from a parental leave program. NBER Working Paper No. 23174, National Bureau of Economic Research.

Fryer Jr., R. G. and S. D. Levitt (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics 86*(2), 447–464.

Fuß, D., T. Gnambs, K. Lockl, and M. Attig (2016). *Competence data in NEPS: overview of measures and variable naming conventions (Starting Cohorts 1 to 6)*. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Galiani, S., M. A. Rossi, and E. Schargrodsky (2011). Conscription and crime: evidence from the argentine draft lottery. *American Economic Journal: Applied Economics 3*(2), 119–136.

Garces, Eliana, T. D. and J. Currie (2002). Longer-term effects of Head Start. *American Economic Review 92*(4), 999–1012.

Gathmann, C. and N. Keller (2018). Access to citizenship and the economic assimilation of immigrants. *Economic Journal 128*(616), 3141–3181.

Gathmann, C., N. Keller, and O. Monscheuer (2019). Citizenship and social integration. Technical report, University of Heidelberg.

Gathmann, C. and B. Sass (2018). Taxing childcare: effects on childcare choices, family labor supply and children. *Journal of Labor Economics 36*(3), 665–709.

Gehrer, K., S. Zimmermann, C. Artelt, and S. Weinert (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online 5*(2), 50–79.

Gelbach, J. B. (2002). Public schooling for young children and maternal labor supply. *American Economic Review 92*(1), 307–322.

Goerres, A. and M. Tepe (2013). Für die Kleinen ist uns nichts zu teuer? Kindergartengebühren und ihre Determinanten in Deutschlands bevölkerungsreichsten Städten zwis-

chen 2007 und 2012. *dms - Der moderne Staat, Zeitschrift für Public Policy, Recht und Management 6*, 169–190.

Gonzalez, L. (2013). The effect of a universal child benefit on conceptions, abortions, and early maternal labor supply. *American Economic Journal: Economic Policy 5*(3), 160–188.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of Child Psychology and Psychiatry 38(5)*, 581–6.

Gormley, W. T. J. and T. Gayer (2005). Promoting school readiness in Oklahoma: an evaluation of Tulsa's pre-k program. *Journal Human Resources 40*, 533–558.

Goux, D. and E. Maurin (2010). Public school availability for two year-olds and mothers' labour supply. *Labour Economics 17*, 951–962.

Grenet, J., R. A. Hart, and J. E. Roberts (2011). Above and beyond the call. Long-term real earnings effects of british male military conscription in the post-war years. *Labour Economics 18*(2), 194–204.

Gruber, J. and S. A. Kleiner (2012). Do strikes kill? Evidence from new york state. *American Economic Journal: Economic Policy 4*(1), 127–157.

Hahn, I., K. Schöps, S. Rönnebeck, M. Martensen, S. Hansen, S. Saß, I. M. Dalehefte, and M. Prenzel (2013). Assessing scientific literacy over the lifespan – A description of the NEPS science framework and the test development. *Journal for Educational Research Online 5*(2), 110.

Havnes, T. and M. Mogstad (2011a). Money for nothing? Universal child care and maternal employment. *Journal of Public Economics 95*(11-12), 1455–1465.

Havnes, T. and M. Mogstad (2011b). No child left behind: universal child care and children's long- run outcomes. *American Economic Journal: Economic Policy 3*, 97–129.

Hecker, K., A. Südkamp, C. Leser, and S. Weinert (2015). Entwicklung eines Tests zur Erfassung von Hörverstehen auf Textebene bei Schülerinnen und Schülern der Klassenstufe 9. NEPS Working Paper No. 53, Leibniz Institut für Bildungsverläufe e.V.

Heckman, J. J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources 32*(3), 441–462.

Heckman, J. J. (2007). The economics, technology, and neuroscience of human capability formation. *Proceedings of the National Academy of Sciences 104*(33), 13250–13255.

Heckman, J. J., S. Urzua, and E. J. Vytlacil (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics 88*(3), 389–432.

Heckman, J. J. and E. J. Vytlacil (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences 96*(8), 4730–4734.

Heckman, J. J. and E. J. Vytlacil (2001). Policy-relevant treatment effects. *American Economic Review P&P 91*(2), 107–111.

Heckman, J. J. and E. J. Vytlacil (2005). Structural equations, treatment effects and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J. and E. J. Vytlacil (2007a). Econometric evaluation of social programs,

part I: causal models, structural models and econometric policy evaluation. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6, Part B, Chapter 70, pp. 4779–4874. Amsterdam: Elsevier.

Heckman, J. J. and E. J. Vytlacil (2007b). 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 environments. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6, Part B, Chapter 71, pp. 4875–5143. Amsterdam: Elsevier.

Herr, A. (2008). Cost and technical efficiency of German hospitals: does ownership matter? *Health Economics 17*(9), 1057–1071.

Herr, A., H. Schmitz, and B. Augurzky (2011). Profit efficiency and ownership of german hospitals. *Health Economics 20*(6), 660–674.

Ho, V. and B. H. Hamilton (2000). Hospital mergers and acquisitions: does market consolidation harm patients? *Journal of Health Economics 19*(5), 767–791.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Imbens, G. and W. V. D. Klaauw (1995). Evaluating the cost of conscription in the Netherlands. *Journal of Business & Economic Statistics 13*(2), 207–215.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature 47*(1), 5–86.

International ICT Literacy Panel (2002). Digital transformation: a framework for ICT literacy. *ETS Report of the International ICT Literacy Panel*.

James-Burdumy, S. (2005). The effect of maternal labor force participation on child development. *Journal of Labor Ecoomics 23*(1), 177–211.

Jensen, G. A. and M. A. Morrisey (1986). Medical staff specialty mix and hospital production. *Journal of Health Economics 5*(3), 253–276.

Kamhöfer, D. A., H. Schmitz, and M. Westphal (2018, 02). Heterogeneity in marginal non-monetary returns to higher education. *Journal of the European Economic Association 17*(1), 205–244.

Kaufmann, K. M. (2014). Understanding the income gradient in college attendance in Mexico: the role of heterogeneity in expected returns. *Quantitative Economics 5*(3), 583–630.

Kline, P. and C. R. Walters (2016). Evaluating public programs with close substitutes: the case of Head Start. *Quarterly Journal of Economics 131*(4), 1795–1848.

Kottelenberg, M. and S. Lehrer (2017). Targeted or universal coverage? Assessing heterogeneity in the effects of universal childcare. *Journal of Labor Economics 35*(3), 609–653.

Kühnle, D. and M. Oberfichtner (2017). Does early child care attendance influence children's cognitive and non-cognitive skill development? IZA Working Paper No. 10661, Institute of Labor Economics (IZA).

Lalonde, R. and R. Topel (1997). Economic impact of international migration and the

economic performance of migrants. In M. Rosenzweig and O. Stark (Eds.), *Handbook of Population and Family Economics*, Volume 1B, pp. 799–850. Amsterdam: Elsevier Science.

Lefebvre, P. and P. Merrigan (2008). Child-care policy and the labor supply of mothers with young children: a natural experiment from Canada. *Journal of Labor Economics 26*(3), 519–548.

Leu, H. R. and M. Schilling (2008). *Zahlenspiegel 2007*, Volume 1, Chapter 9, pp. 219–232. Munich: Deutsches Jugendinstitut e.V.

Lin, H. (2014). Revisiting the relationship between nurse staffing and quality of care in nursing homes: an instrumental variables approach. *Journal of Health Economics 37*(Supplement C), 13 – 24.

Love, J. M., E. Eliason, C. Ross, H. Raikes, J. Constantine, and K. Boller (2005). The effectiveness of early Head Start for 3-year-old children and their parents: lessons for policy and programs. *Developmental Psychology 41*(6), 885–901.

Ludwig, J. and D. L. Miller (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics 122*(1), 159–208.

Lundin, D., E. Mörk, and B. Öckert (2008). How far can reduced childcare prices push female labour supply? *Labour Economics 15*, 647–659.

Maestas, N., K. J. Mullen, and A. Strand (2013). Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. *American Economic Review 103*(5), 1797–1829.

Maurin, E. and T. Xenogiani (2007). Demand for education and labor market outcomes lessons from the abolition of compulsory conscription in France. *Journal of Human Resources 42*(4), 795–819.

Miller, D. L., C. A. Cameron, and J. B. Gelbach (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics 90*, 414–27.

Monscheuer, O. (2019). National identity and the integration of second-generation immigrants. Technical report, University of Heidelberg.

Needleman, J., P. Buerhaus, S. Mattke, M. Stewart, and K. Zelevinsky (2002). Nurse-staffing levels and the quality of care in hospitals. *New England Journal of Medicine 346*(22), 1715–1722.

Neumann, I., C. Duchhardt, M. Grüßing, A. Heinze, E. Knopp, and T. Ehmke (2013). Modeling and assessing mathematical competence over the lifespan. *Journal for Educational Research Online 5*(2), 80–109.

Neyman, J. and K. Iwaszkiewicz (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society 2*(2), 107–180.

Nollenberger, N. and N. Rodriguez-Planas (2015). Full-time universal childcare in a context of low maternal employment: Quasi-experimental evidence from Spain. *Labour Economics 36*, 124–136.

Nolting, H., K. Zich, B. Deckenbach, A. Gottberg, K. Lottmann, D. Klemperer, M. Grote Westrick, and U. Schwenk (2011). *Faktencheck Gesundheit - Regionale Unterschiede in der Gesundheitsversorgung*. Gütersloh: Bertelsmann Stiftung.

Nybom, M. (2017). The distribution of lifetime earnings returns to college. *Journal of Labor Economics 35*(4), 903–952.

OECD (2011). *Naturalisation: A Passport for the Better Integration of Immigrants?* Paris: OECD Publishing.

OECD (2013, 2). How do early childhood education and care (ECEC) policies, systems and quality vary across OECD countries? *Education Indicators in Focus*, No. 11. Paris: OECD Publishing.

OECD (2016). *Health Workforce Policies in OECD Countries - Right Jobs, Right Skills, Right Places*. OECD Health Policy Studies. Paris: OECD Publishing.

OECD (2017). *Starting Strong 2017: Key OECD Indicators on Early Childhood Education and Care*. Paris: OECD Publishing.

OECD (2018a). *International Migration Outlook 2018*. Paris: OECD Publishing.

OECD (2018b). *The Resilience of Students with an Immigrant Background - Factors that Shape Well-being*. OECD Reviews of Migrant Education. Paris: OECD Publishing.

Oi, W. Y. (1967). The economic cost of the draft. *The American Economic Review 57*(2), 39–62.

Propper, C., S. Burgess, and K. Green (2004). Does competition between hospitals improve the quality of care? Hospital death rates and the NHS internal market. *Journal of Public Economics 88*(7-8), 1247–1272.

Propper, C. and J. V. Reenen (2010). Can pay regulation kill? Panel data evidence on the effect of labor markets on hospital performance. *Journal of Political Economy 118*(2), 222–273.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica 56*, 931–954.

Robson, A. L. (2002). Critical and sensitive periods. In N. J. Salkind (Ed.), *Child Development (Macmillan Psychology Series)*, Volume 1, pp. 101–103. New York: Macmillan.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688.

Sajons, C. (2016). Does granting citizenship to immigrant children affect family outmigration? *Journal of Population Economics 29*(2), 395–420.

Sandner, M. and T. Jungmann (2016). How much can we trust maternal ratings of early child development in disadvantaged samples? *Economics Letters 141*, 73–76.

Schlosser, A. (2006). Public pre-school and the labor supply of Arab mothers: evidence from a natural experiment. *The Economic Quarterly Vol. 53*(No. 3), 517–553.

Schmitz, H. and H. Tauchmann (2012). Factor substitution in hospitals: a DEA based approach, *SFB 823 Discussion Paper No. 41*.

Schwarz, N. (1999). Self-reports. How the questions shape the answers. *American Psychologist 54*(2), 93–105.

Senkbeil, M., J. M. Ihme, and J. Wittwer (2013). The Test of Technological and Information Literacy (TILT) in the National Educational Panel Study: development, empirical testing, and evidence for validity. *Journal for Educational Research Online 5*(2), 139–161.

Shachar, A., R. Bauboeck, I. Bloemraad, and M. Vink (Eds.) (2017). *The Oxford Hand-*

book of Citizenship. Oxford, UK: Oxford University Press.

Siegler, R., J. Saffran, N. Eisenberg, J. Deloache, and E. Gershoff (2017). *How Children Develop* (5th ed.). New York: Worth Publishers.

SOEP (2017). Socio-Economic Panel, Data for Years 1984-2016, Version 33, SOEP 2017.

Sohns, A. (2009). Pädagogische Konzepte in Kindertagesstätten. In R. Stein and D. Orthmann Bless (Eds.), *Basiswissen Sonderpädagogik*. Baltmannsweiler: Schneider-Verlag.

Sparrow, S. S., D. V. Cicchetti, and D. A. Balla (2005). *Vineland Adaptive Behavior Scales: Survey Forms Manual*. Circle Pines, MN: American Guidance Service.

Steinhardt, M. F. (2012). Does citizenship matter? The economic impact of naturalizations in Germany. *Labour Economics 19*(6), 813–823.

Steinmann, L. and P. Zweifel (2003). On the (in)efficiency of Swiss hospitals. *Applied Economics 35*(3), 361–370.

Sweetman, A. and J. C. van Ours (2015). Immigration: what about the children and grandchildren? In B. R. Chiswick and P. W. Miller (Eds.), *Handbook of the Economics of International Migration*, Volume 1b, Chapter 21, pp. 1141–1193. Amsterdam: Elsevier.

Tong, P. K. (2011). The effects of California's minimum nurse staffing laws on nurse labor and patient mortality in skilled nursing facilities. *Health Economics 20*(7), 802–816.

Treutler, C. M. and C. C. Epkins (2003). Are discrepancies among child, mother, and father reports on children's behavior related to parents' psychological symptoms and aspects of parent-child relationships? *Journal of Abnormal Child Psychology 31*(1), 13–27.

Van Der Klaauw, B. (2014). From micro data to causality: forty years of empirical labor economics. *Labour Economics 30*, 88–97.

Walters, C. (2015). Inputs in the production of early childhood human capital: evidence from Head Start. *American Economic Journal: Applied Economics 7*(4), 76–102.

Weinert, S., C. Artelt, M. Prenzel, M. Senkbeil, T. Ehmke, and C. H. Carstensen (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft 14*(2), 67–86.

Wissenschaftliches Institut der AOK (2007). *Qualitätssicherung der stationären Versorgung mit Routinedaten (QSR)* (1 ed.). Bonn: Wissenschaftliches Institut der AOK.

Worbs, S. (2014). *Bürger auf Zeit. Die Wahl der Staatsangehörigkeit im Kontext der deutschen Optionsregelung*. Ph. D. thesis, University of Education Schwäbisch Gmünd, Nuremberg.