Essays in Education and Health Economics

Ilka Gerhardts

Dissertation
Munich, 2022

Essays in Education and Health Economics

Inaugural-Dissertation
Zur Erlangung des Grades Doctor oeconomiae publicae (Dr. oec. publ.)
an der Ludwig–Maximilians–Universität München

2022

vorgelegt von
Ilka Gerhardts

Referent: Prof. Dr. Uwe Sunde
Korreferent: Prof. Dr. Joachim Winter
Promotionsabschlussberatung: 02. Februar 2022

Datum der mündlichen Prüfung: 05. Juli 2021
Namen der Berichterstatter: Prof. Dr. Uwe Sunde
Prof. Dr. Joachim Winter
Prof. Dr. Ludger Wößmann

Contents

List of Figures

List of Tables

Acknowledgments

Happily, I grasp the opportunity for a written thank-you to people that gave my life countless positive twists.

First and foremost, I thank Uwe Sunde, my principal advisor, who reshaped my professional life beyond words. He encouraged me to pursue an academic career and cares about our joint and my other projects alike. I could not wish for a better mentor. My gratitude also goes to Joachim Winter, my second advisor, for providing valuable, hands-on, and constructive feedback. I appreciate the new perspectives Ludger Wößmann added to my research. Carsten Eckel owns my gratitude in many ways. As his administrative assistant, I enjoyed a *tierischen* sense of humor.[1]

I am indebted to my colleagues at the Chair for Population Economics, each of them being quite a character. Very special thanks go to Lukas Buchheim, my junior mentor, whose humane pedagogical approach follows in the tracks of Uwe Sunde. I thank Sabine Njoku, always cheerful and minimizing administrative challenges, just as Susan Fay did before. I admire Ursula Baumann for rigorously enforcing nature-friendly exam corrections, thereby being ahead of her time by decades. A huge hug goes to Miri(am) Breckner, one of my best friends.

As a member of MGSE, and EBE in particular, I was guaranteed an interactive and productive research environment, in which Florian Engelmeier plays a key role. I am genuinely happy the doctoral program connected me to Zhaoxin Pu, another best friend of mine, as well as to Patrick Reich, my sportiest friend, Nathalie Obergruber, and Annika Bergbauer. Let me highlight that Anna Gumpert, Amelie Schiprowski, Sven Resnjanski, Mira Fischer, and Markus Nagler spend much time sharing tips on making the PhD a worthwhile experience. I thank Andreas Steinmayr and Manuel Voigt for statistical insights and my co-author Larissa Zierow for employing me as a programming RA. In turn, the RAs David Greisberger, Rahela Udiljak, Alexander Wilder Quispe, and Simon Velten supported my projects smartly and diligently.

I am indebted to the Joachim-Herz-Stiftung and the DFG for research funding. The

[1] Carsten is my dog's personal photographer.

Fulbright Foundation enabled a research stay at UCSB, where I was warmly received as a Broom Center visitor. USCB's doctoral students as well as my flatmate, Rose Jei Gion, were incredibly helpful in settling in. Douglas Steigerwald (and his Econometrics reading group) repeatedly provided granular feedback for my work. Last but foremost, Shelly Lundberg, my generous host, kindly offered weekly counseling meetings, at one of which a chapter of this thesis was born. Dimitra Bon, assigned by the IQTIG data institute to execute that project, handled the extensive code with dedication, even running the most demanding snippets overnight. Which reminds me of Felix Grein's patience debugging IT nuisances of any type.

Remembering my student exchange to Spain's UNIZAR, special thanks belong to Maribel Ayuda Bosque, my professor of Econometrics, who fueled my interest in empirical work. At this point, I also refer to my *Onkel Manfred*, our family's globe trotter, who offered insider advice for my first longer stay abroad. During the master's at LMU, I was lucky to meet Iryna Kozak Kozak. The joint studies in the research track mark our friendship till today. Bernhard Seimel and my other colleagues from the LIGA Bank supported my part-time working schemes compatible with master's studies, after Sebastian Braun, Georg Baumgartner, and my *Großonkel Quentin* had facilitated my transition from the FernUni Hagen to LMU. Patrizia Dilly and Petra Kallenbach ensured survival in a yet earlier, very different working environment. Sr. Anna Flasza and Rebecca Eibl made me feel at home in Munich within a few weeks of moving here.

I feel that my family and my friends relate in many ways to my research projects. On the one hand, by working in education production, like both my parents and my sisters, as well as Catharina Rottinger. On the other hand, I am intrigued by the broad range of mothers who kindly discussed my obstetric-care project from a less theoretical perspective than myself, among others my own mother, my office mate Johanna Rude, my sister Nina Steinhausen, and Eva Bühren, my early-childhood friend, now mother to four boys and a giant dog. I love the walks with Anna Kuhn, as much for relaxation as for academic insights. I am indebted to Annabel Fleschutz, my favorite *leader* in ballroom dancing, for correcting parts of this thesis, and for inspiring talks, walks, and Yoga flows. My connection to Yoga I owe to Sarah Leinweber, with whom I shared happy times at high school. Whenever Yoga was no help, Susanne Wessloy and Samuel Bonorden, my physicians, artfully resolved all aches and pains from hours of sitting.

I appreciate the all-round support of Francisco de la Mata, who is (nearly) as funny as Carsten. I have been warned to make no high and mighty love declaration printed here forever. Therefore, all I say is, not being with you, *Frani*, I would stay single. Which admittedly has to do with your relation to tender-hearted Paqui Casas y Julián de la Mata inviting me every summer to their Spanish oasis of pomegranates and Paso Dobles,

boosting my energy for the winter term.

Turning to my German family, there is *Elmi* - a great companion of my early childhood, enabling my mother to work full-time with three kids. Oh, *Mami*, from my perspective today, you cut short on your leisure time (too) much. As a child, I believed the tremendous workload was a game, I just loved stamping little appreciative *Tierchen* on the exams you corrected in the evenings (I could not even read back then but felt very important in helping you). Now retired, you taught yourself more Spanish than I could speak after school. I think you are incredible. *Papi*, a lifetime of (brain) jogging and still sportier than I am. When you wrote a book I hammered along on the ancient typewriter making more noise than sense. This was our first project and more are to come. In the Hochsauerland, you two created a green refugium for us, which is the true home, no matter where we live with our first residence. *Nina*, my oldest sister and family's artist, every time we visit, you integrate me and my dog into your busy life - despite the private zoo and (school/ yoga) teaching load. *Lara*, my second-oldest sister, we explored all the wilderness of the Sauerland's deep forests and with your magic brains, you created a fantastic childhood world that still lives in me. And finally, *Meersi* and *Mowgli*, – one stroke of your *lomo* melts the stress away. Cómo se mide aquel regalo de la naturaleza? Veintidos kilitos de felicidad...

Source: With the courtesy of Nina Steinhausen (née Gerhardts)

All people mentioned here have their (more or less direct) fair share in the achievements of my thesis. All errors are my own.

Introduction

"Human rights are universal and inalienable. Human Rights standards – to food, health, education, [...] – are also interrelated. The improvement of one right facilitates advancement of the others. Likewise, the deprivation of one right adversely affects the others." (WHO, December 2017)

Health and education are essential for the well-being of the individual and society as a whole. However, large and persistent heterogeneity around the globe has puzzled economists for many years now: Where does inequality come from, when did it start, and how can it be resolved (Spolaore and Wacziarg, 2013)? Identifying the key role of health and education in fighting inequality (Currie, 2009) was a milestone that readily raised new questions, among the most important ones: When is the optimal moment to invest in either one? While earlier work like Coleman (1968) focused on schooling and adult education, Heckman (2011) documents that human capital benefits of early-childhood education outweigh later-life education by far. Persson and Rossin-Slater (2018) suggest that health-related inequality can be traced even further back, up to the prenatal period.

This thesis provides scientific evidence relevant to current policy debates about the quality of public education and health provision in Germany. Despite Germany's relative wealth in global comparison, anecdotal evidence points to deficiencies in both, the health and the education sector. Addressing these issues, the thesis moves from research on secondary education, over primary education, to economic implications of birth modes. All analyses exploit quasi-experimental set-ups with microeconometric methods applied to (mainly) large-scale registry data.

Education in Germany has seen a series of changes since the 1950s. Many states rolled out major reforms through which the school entry age was standardized, denominational schools turned into common schools, a flexible school entrance phase was introduced, etc. Taken together, schooling reforms changed classroom composition by various dimensions, especially age, culture, and gender. Hattie (2002) recaps the large but predominantly observational literature riddled with inconsistent findings. In particular, it remains unclear which peer composition dimension is the most relevant one and to which extent they interact: Is there is a gender-specific component to heterogeneity in terms of age or culture? This thesis explores quasi-experimental changes in classroom peer composition with and without a pedagogical framework. Disentangling a multitude of mechanisms, multiple significant effects on an individual's success are confirmed and some ambiguities from the literature are reconciled.[2]

As for health, in German hospitals the rate of induced labor at childbirth has risen close to 22.5% in 2019 (DGGG, 2020a), nearly doubling the rate of 1985 (Schwarz, 2008). Staff shortages alleviated through scheduling labor and other birth interventions are openly debated.[3] However, causal evidence on the impact of non-medically motivated induction is so scarce that major obstetric induction guidelines rely on observational evidence or RCTs of disputed internal and external validity.[4] Because of data limitations, most quasi-experimental studies report health effects of immediate induction relative to either postponed induction or any birth mode after some waiting period.[5] This misses out on the potentially most important impact of induction (alone or in combination with surgical interventions) vs. unassisted labor. To address this issue, this thesis introduces a new framework to assess the impact of physician-induced demand for labor induction. It quantifies both, the detrimental effects on a mother's and her neonate's health as well as the extent to which the health effects bounce back on a hospital's staff capacity through patient monitoring requirements.

The first two chapters focus on educational attainment driven by heterogeneous class-

[2] (Quasi-)experimental studies on age heterogeneity effects by, e.g., Checchi and De Paola (2018b) and Sims (2010) document negative effects. Leuven and Rønning (2016) confirm negative effects on relatively older students but find positive impacts on younger students, both of which stronger for girls. Targeting gender heterogeneity directly, Whitmore (2005) finds positive effects of a higher girls' share. Lee et al. (2015) and Gneezy et al. (2003) report positive co-education effects for boys and (weakly) negative ones for girls. Booth and Nolen (2012) support negative impacts on girls, Hill (2015) on both genders. Turning on cultural heterogeneity, Lavy and Schlosser (2011) find positive effects of gender heterogeneity within Jewish schools. Figlio and Stone (1999) report overall positive effects of common compared to Catholic schools.

[3] For anecdotal evidence, see articles headlined *'Die pauschale Geburt'* (FLatrate Birth. Own translation. Deutsches Ärzteblatt, June 2014) and *'Geburtshilfe: Wenn die Aufnahme in den Kreißsaal vom Zufall abhängt'* (Obstetric care: When admission to the labor room is left to chance. Own translation. Süddeutsche Zeitung, September 2017). Scientific evidence linking birth interventions to a hospital's organization goes back to (Brown, 1996).

[4] In response to a large-scale RCT hampered by methodological shortcomings (Carmichael and Snowden, 2019), the DGGG (2020a) and the ACOG (2021) allow offering induction routinely to healthy first-time mothers at term. In the same guideline, the DGGG (2020a) states that the health sectors' total costs arising from routine inductions are still unknown. Notably, the WHO (2018), building on relatively stronger scientific evidence (Tsakiridis et al., 2020), stuck with a more conservative approach.

[5] Buckles and Guldi (2017), Jürges (2017), Lynch et al. (2019), Gans and Leigh (2008), and Schulkind and Shapiro (2014a) find predominantly negative or zero health effects of induction.

room composition in terms of denomination, age, and gender.[6] Chapter 1 studies denominational schools as an important provider of education in many countries around the world. Due to their focus, these schools often operate with multigrade classes, in which more than one age cohort is taught in one classroom. Multigrade classes are a cost-effective way to provide education and play a crucial role in education policy in the context of demographic change. This chapter presents estimates of the causal effect of attending denominational schools with multigrade classes on schooling and short-run labor market outcomes. The analysis combines administrative records of schools with comprehensive population census data, and exploits the abolition of denominational schools in the Saarland, a German state, in 1969, for identification of the effect.[7] The findings document significantly detrimental effects on final grade attainment and labor market participation. Notably, the negative impact is most pronounced in the outcomes of girls.

Chapter 2 provides novel evidence on the causal effect of multigrade teaching in primary schools on literacy skills by the end of primary school. The analysis is based on student test score data of more than 68'000 fourth-graders and exploits the staggered introduction of policies targeted at making entry to primary schools more flexible across German states between 2001 and 2016 for identification. The results from a difference-in-differences design document that attending multigrade classes had negative effects on reading test scores and German grades. These negative effects are again centered on girls.

Chapter 3 studies maternal and neonatal health, including its rebound on staff capacity, in the face of physician-induced demand at childbirth. It documents alarming causal evidence on the negative consequences of non-medically indicated induced labor (and surgical interventions), first and foremost the increased incidence of severe perineal tearing, lower APGAR scores, and (following surgical interventions) sizably prolonged postnatal stays.

Birth interventions are common practice in OECD countries, especially in German hospitals, where profit-oriented reimbursement schemes and acute staff shortages incentivize intervention for organizational relief. The identification strategy exploits exogenous variation in staff shortages and physician-specific intervention preferences, both of which are documented in two years of nationwide comprehensive hospital records. The design overcomes non-random and interdependent assignment of induced labor with/-out non-emergency c-sections, and vaginal operations, thereby distinguishing successful inductions and those leading to surgical delivery. The disaggregated insights from decomposing the

[6] They are based on joint work with Uwe Sunde and Larissa Zierow.

[7] Saarland's reform impacted 95% of schools. However, denomination schools still exist in some states, nowadays criticized for exploiting their Christian label to discriminate against foreign, mostly Muslim, children (Spiegel, August 2009).

most relevant (single and combined) treatment effects confirm that even simple multi-treatment models are preferable over single-treatment estimation.

Finally, ongoing work by Gerhardts (2024) explores possibly heterogeneous effects across maternal and hospital strata, most importantly smaller vs. low-quality hospitals, as well as relatively older, less educated, and single mothers, thereby working towards the causal link between health and education suggested by Heckman (2007) and Currie (2009).

Chapter 1

Multigrade Classes and Returns to Education: Evidence from the Abolition of Denominational Schools

1.1 Introduction

Many schools are operated on a basis of multigrade classes. Multigrade teaching represents a cost-effective way of providing children with education in the context of limited resources. In fact, in large parts of the world schools with multigrade classes, often run by different religious denominations, represent the typical way of teaching children. Around the globe, approximately one third of all classes across all countries, including some of the more developed countries, are multigrade classes (2005 UNESCO Agenda for Educational Planning).

Multigrade classes have recently become a principal adjustment device for enrollment fluctuations also in many parts of Europe where demographic aging puts increasing pressure on class sizes. Warnings have been raised regarding the potentially detrimental effects of teaching students of different ages and maturity within the same room. At the same time, teaching several cohorts in one classroom has been suggested to have advantageous pedagogical side effects by providing more intense interactions between students of different ages that foster student-based learning.

Historically, many schools were restricted to particular religious denominations, which led to a restriction of student numbers, and consequently multigrade teaching, as the result of religious segregation. With denominational affiliation losing importance, this

This chapter is based on joint work with Uwe Sunde and Larissa Zierow.

led to the abolition of denominational schools in many parts of Europe. At the same time, this abolition lifted size restriction and led to the abolition of multigrade classes. Mixed empirical evidence regarding the effects of abolishing denominational schools with multigrade classes on subsequent outcomes continues to fuel heated debates regarding the appropriate school organization.

This paper investigates the impact of the abolition of denominational schools with predominantly multigrade teaching on the long-term returns to education. The identification strategy exploits the natural experiment of a large-scale reform that led to the abolition of denominational schools in the Saarland, a state in Germany, in 1969. Prior to the reform, more than 95% of primary and lower secondary schools were church-maintained. In scarcely populated regions, the strict tracking by religious denomination imposed severe restrictions on the allocation of students. As a consequence, schools were relatively small, implying that students of different ages and skills were taught within the same classroom, i.e. in multigrade classes. The abolition of denominational schools in 1969 led to the dissolution of hundreds of these rural multigrade schools within less than a year. The remaining schools obtained a single-grade structure, similar to the larger schools in more urban environments.

The identification approach exploits differential treatment exposure of students depending on how many students of the same birth cohort have the same denomination. In more rural municipalities, multigrade teaching in denominational schools was the norm prior to 1969, but not afterwards. By contrast, in more urban municipalities multigrade teaching in denominational schools was not necessary due to higher student numbers. To estimate the effects of the reform on schooling and labor market outcomes we use an enhanced differences-in-differences approach.

By exploring the heterogeneity of the effects across gender, the evidence also provides new insights into the roots of gender inequality. In particular, the large-scale natural experiment enables insights into the socialization mechanisms at school that might lead to gender differences in labor market participation and occupational choice later on in life.

The empirical analysis is based on a unique combination of administrative records and comprehensive population census data. The dataset has been collected and digitized specifically for this research project, which to our knowledge is the first to exploit the abolition of denominational schools as a natural experiment in this context. Using municipality codes and schools' denominations, we are able to link individual-level census data on virtually all of Saarland's households in 1970 and 1987 to a comprehensive schools' index that comprises more than 7,500 school-year observations on a municipality-denomination-level. The availability of a wide range of schooling covariates allows us to control for channels like class size, school size, school consolidation, gender composition, etc. that might confound the multigrade effects.

The empirical results suggest that the abolition of multigrade classes had positive effects on final grade attainment and labor market participation. While all students profited from the abolition of denominational schools in terms of the higher grade attainment and a greater likelihood to become a white-collar worker, the effect is notably stronger for girls. The abolition of denominational schools in municipalities where multigrade teaching was the norm before 1969 led to an increase in the number of girls who attained a higher educational degree and a decrease in the number of girls becoming housewives. The results therefore suggest an interplay of gender socialization and the mode of teaching in terms of multigrade classes on subsequent outcomes.

The question how denominational schools with multigrade classes affect students' outcomes touches upon several research strands related to class composition, educational infrastructure, peer and tracking studies. Our empirical approach contributes to the literature in several ways. First, the natural experiment of the sudden abolition of denominational schools allows for a credible identification of the causal impact of denominational schools with multigrade classes, whereas many existing studies suffer from insufficient randomization which renders identification problematic (mainly because of self-selection). Second, we present effects that are placed in a Western European society. Many studies on multigrade classes with credible identification (due to controlled randomization) have been conducted mainly in developing countries, at the cost of limited external validity for more developed countries. Moreover, recent studies on multigrade teaching with credible identification focus on short-term educational outcomes. Third, the high-quality dataset covering virtually the complete population of our region of study minimizes selection and response biases and affords statistical power whereas existing research mostly relies on evidence from small samples. Fourth, provided with large-scale evidence, we are able to link gender mechanisms at school not only to final grade attainment but also to labor market participation and occupational choice. Our analysis thereby extends earlier work that mainly focused on the gender specific effect of class composition on schooling outcomes. Overall, our results are in line with the findings of earlier studies that suggest rather negative effects of multigrade classes.

The remaining part of the paper is structured as follows. Section 1.2 gives an overview of the existing literature on class composition. Section 1.3 describes the institutional background. Section 1.4 presents the identification strategy, followed by a compact presentation of the data in Section 1.5. Section 1.6 presents the empirical results, discusses robustness with respect to sensitivity checks and shows the results of the subgroup analysis. Section 1.7 concludes.

1.2 Literature Review

Multigrade classes[1] produce multiple forms of *peer effects*. Peer effects are central aspects of education research. They have been modeled as inputs to the education production function ever since Coleman (1968) made them popular, among others by Iversen and Bonesrønning (2015); Jones (2013). There exists relatively less research on peer effects of class composition than, e.g., on class size (Jones, 2013), but the absolute number of class composition studies is still vast. Many of those have been criticized for low methodological quality, however, as detailed in Lindström and Lindahl (2011) or Mason and Burns (1996). In general, a variety of peer effects can arise in a system of multigrade classrooms which has been touched upon as follows.

Between-student spillovers may be positive if more knowledgeable, skilled or able classmates serve as natural role models (Duflo et al., 2011; Hanushek et al., 2003). Practical relevance of peer collaboration, however, is told to be rather limited (Hattie, 2002). There is also evidence that peer effects are rendered negative if age gaps arise due to grade repeating and redshirting which is often the case in developing countries (Lavy et al., 2012; Jones, 2013).

Finally, peer effects among teachers in the sense of shared experiences have been mentioned in the multigrade context. The probability of beneficial spillovers prerequisites at least two teachers per school and is likely to increase in larger teaching staff which puts rural schools at a disadvantage (McEwan, 2008).

Besides peer effects, also effects of (no) adjustments of teacher training, curricula, materials and incentives need to be reconsidered upon collapsing the grade level structure. Traditional teacher colleges prepare single-grade teaching although multigrade teaching is strategically more demanding and stressful (Mason and Burns, 1996; Russell et al., 1998). Therefore, it is likely that multigrade schools have negative effects on students if the pedagogical infrastructure is not adapted to multigrade teaching.

Current research on multigrade classes is frequently located in developing countries. See Little (2001) or McEwan (2008) for overviews in Africa, Asia and Latin America respectively. While some randomized control studies conducted in these countries convince by providing internal validity, their external validity is rarely given.[2] First, there are several institutional deficiencies that make it difficult to compare the examined multigrade

[1] Multigrade classes, as opposed to single-grade classes (Veenman, 1995), do not sort students by age and skill. Furthermore, they are created out of some necessity, not pedgogical purpose, as other types of combination classes are.

[2] Not only randomized control studies deliver evidence for multigrade effects in developing countries. Jones (2013) relies on an IV strategy to circumvent selection issues. He presents strongly negative effects by African overage-for-grade peers thus being supportive of Lavy et al. (2012).

settings to each other. For example, in some cases the mixed grade levels are not even adjacent (Mulkeen and Higgings, 2009) which increases the heterogeneity in the classroom substantially.[3] Second, unsafe school ways complicate school attendance asymmetrically for girls which changes the classroom gender distribution (Mulkeen and Higgings, 2009). Third, grade attainment may not mean anything regarding knowledge and skills (Jones, 2013). Due to this range of peculiarities in developing countries estimation of the effects of multigrade classrooms is challenging even to (quasi-)experimental designs that are good practice in the sense of Angrist (2004).[4]

Even though the major part of research on multigrade classes studies multigrade settings in development countries multigrade classrooms are also prevalent in more developed countries. Contemporaneously, multigrade classes make up one third of all classes on earth, and even in countries like Finland, the Netherlands, India, Peru, Sri Lanka and Pakistan multigrade predominate single-grade classes (Mulkeen and Higgings, 2009).

Existing studies on multigrade classes that were (mostly) conducted in industrialized countries up to 1995 are summarized in a meta-analysis by Veenman (1995). He concludes there are no significant effects on cognitive and/or social-emotional outcomes after averaging over 43 combination class studies meeting his econometric criteria. Apart from being quite outdated today these criteria were already criticized by contemporary scholars Mason and Burns (1996). They point out that Veenman (1995) draws on studies that use non-random samples. They argue that multigrade classes have better teachers and students. By that the group composition in multigrade classrooms biases an actually negative effect of less effective teaching in this setting towards zero.[5]

A rather recent study on combination classes is the one by Lindström and Lindahl (2011). They rely on survey data and compare non-random but observationally equivalent single-grade and mixed-age classes in Sweden. They report a negative impact as sizable as that observed for larger classes in the STAR experiment.[6] Another recent approach

[3] Furthermore, teachers in these countries often undergo very different trainings and the rate of teacher absence is very high. Enrollment is not compulsory but rather an achievement in itself, at any age (Jones, 2013).

[4] Vivalt (2015) establishes the overall limited external validity of impact evaluation studies formally.

[5] Concretely, multigrade teaching is found to cover less curriculum, especially in higher grades. Russell et al. (1998) back up the hypothesis that multigrade teaching is increasingly detrimental beyond basic skill acquirement. Furthermore he finds numeracy skills to suffer more than literacy from a multigrade structure in elementary schooling. To the extent of bias due to peer ability Mason and Burns (1996)'s critic is mitigated by Cullen et al. (2006). They present evidence from US school choice lotteries claiming no significant influence on student attainment by higher peer quality associated with the preferred schools. Their quality indicator measures the difference between (single-grade) classmates' average test scores after winning or loosing the lottery. Insignificance applies uniformly to ability, gender and race strata. It is also robust to all intensities of lottery-induced peer improvement.

[6] In the STAR framework the presence of about six more students reduces test scores of classmates by 4

to estimate effects of multigrade classrooms is presented by Leuven and Rønning (2016). Looking at multigrade schools in Norway they highlight the idea of *perspective-dependent* peer instruments obtaining contrastive signs out of the same data. They find younger students to benefit from having older ones around while older students get worse results when younger ones are around.[7] Leuven and Rønning (2016) conclude seemingly inconsistent evidence to be rooted in researchers' unilateral approaches. Furthermore, they claim to reconcile the literature finding small but significantly positive peer effects conditional on an optimal allocation.[8] Subsequent investigations by Carrell et al. (2013), however, point out limitations of peer group interventions as proposed by Leuven and Rønning (2016) in the face of endogenous subgroup formation. They deliberately allocate weak and strong ability students enabling theoretically the largest possible spillovers. They do not foresee more able students to cut less able ones out of their circle leaving them with even worse academic attainments. Recent work by Checchi and De Paola (2018a) estimate the effect of multigrade classes on the formation of student cognitive and non-cognitive skills exploiting institutional features of the Italian educational system establishing a minimum number of students per class. In a companion paper (Gerhardts et al., 2021b) we provide evidence on the causal effect of multi-grade teaching in primary schools on literacy skills by the end of primary school exploiting the variation i policies across the federal states in Germany.

In view of the existing research on multigrade classes our study contributes to the literature in several ways: Our study focuses on the impact of the multigrade setting in German schools and uses a natural experiment – the sudden abolition of denominational schools – for identification of the causal effect of multigrade schools. By contrast, existing studies like those of Lindström and Lindahl (2011) and Leuven and Rønning (2016) suffer from insufficient randomization and rely on selection-on-observables methods which render causal identification problematic. Furthermore, we present effects of multigrade classes that are placed in a Western European society while those studies on multigrade classes with credible identification have been conducted mainly in developing countries. But, as described above, there are quite a few limitations of the institutional settings in these

percentage points in the first year and 1 additional percentage point in subsequent years (Krueger, 1999).

[7] Concretely, they refer to Sims (2010) deriving negative impacts from measuring exposure to lower grade levels thus taking the perspective of the harmed older students. Along the same pattern Thomas (2012) is expected to find positive peer effects because he considers higher grade levels that are taught together with the treated younger students.

[8] Similarly Duflo et al. (2011) uncover contrastive spillover effects for high and low achievers in Indonesian (single-grade) schools. However, after taking into account lasting consequences of more adequate curricula (detailed in Glewwe et al. (2009)) and teachers' tendency to teach to the top of the class, Duflo et al. (2011) find tracking to be beneficial for all students. Yet another (single-grade) example where curriculum adjustments persistently outweigh peer effects is presented by Cortes and Goodman (2014) looking at US schools.

countries which diminishes the external validity of the findings for industrialized countries. Additionally, we possess a high-quality dataset covering virtually the complete population of our region of study. Thus, we do not have to deal with selection and response biases as much as studies relying on survey data (such as Lindström and Lindahl, 2011).

Another advantage of being provided with large-scale evidence is that we are able to explore the effects of multigrade classrooms not only with respect to final grade attainment (as most existing research is confined to) but also to labor market participation and occupational choice. Extending the multigrade analysis to an interplay of medium-run outcomes (as pioneered in other contexts by Clark and Del Bono, 2016; Greenwood et al., 2016) is new to the literature.

1.3 Institutional Background

This section describes the school reform in the region of our study, the framework of schooling laws, as well as potential confounders, using information from various sources.

Prior to the reform in 1969, almost all *Volksschulen* sorted students by denomination. This allocative restriction created multigrade classes in regions with a low population density. Figure 1.1 provides a first overview of the prevalence of multigrade classes in the Saarland prior to the reform.[9] With few exceptions denominational schools played a role only in the lowest educational track. For a concise description of ability tracking in German schools see Pischke and Wachter (2005).[10]

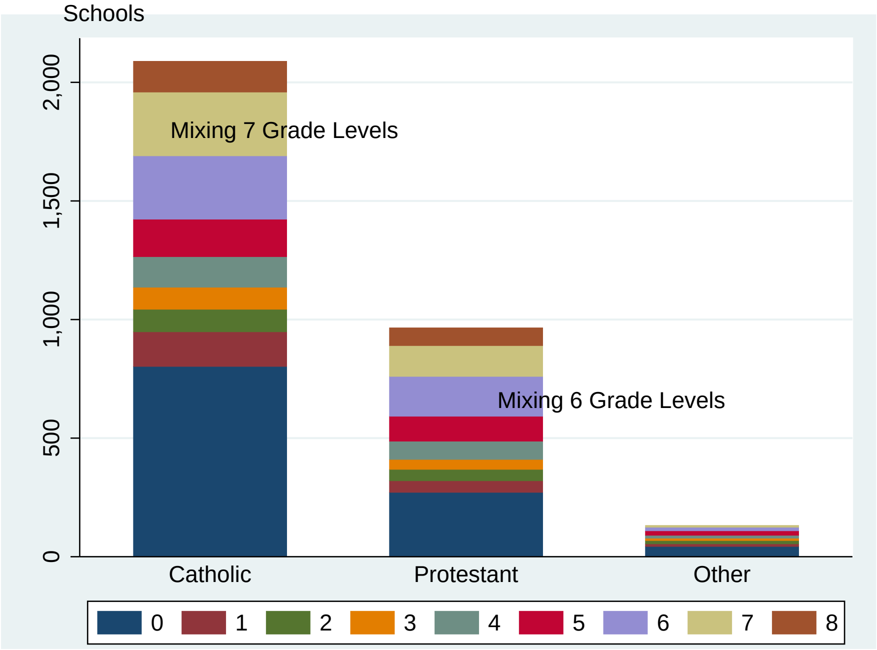
Schools providing primary or lower secondary education were uniformly labeled *Volksschule*, see Figure 1.A.1 in the appendix for a more details on the distribution of school types over time.

Prior to the abolition of denominational schools, the treatment exposure (the prob-

[9] Rural *Volksschulen* create a multigrade setting not supported by pedagogical adjustments. First, the schools' records do not provide any evidence for adjustments. Moreover, albeit this is no rocket-science, there do exist alarming hints about amateurishly adapted teaching practices, available at `http://www.spiegel.de/spiegel/print/d-46265072.html` (01 May 2015). which highlights the comparability problem to mixed-age classes (Mulkeen and Higgings, 2009).

[10] Multigrade classes in remote regions pool children of very different abilities. Do the observed spillovers of our study provide guidance for inclusion of handicapped children as well? This depends on the multigrade school employing a full inclusion policy. Iversen and Bonesrønning (2015) explore spillovers in Norwegian elementary schools where special education happens to be integrated within ordinary classrooms. They find that spillovers interact with the level of special education provided. In Germany the *Volksschule* and special schools are kept apart. After reforming lower secondary education the separation persists (Figure 1.A.1). Thus the insights by Iversen and Bonesrønning (2015) formalize the lack-of-comparability argument forwarded in Veenman (1995) by which he excludes studies on gifted as well as handicapped children from his synthesis.

Figure 1.1: Mixed Grade Levels by Denomination



Notes: This figure shows the prevalence of multigrade teaching prior to the reform in 1969 by denomination. The category 'Other' mainly consists of non-denominational schools. Each color represents the amount of grade levels that were taught together. Red, for instance, shows the number of schools that were teaching 5 grade levels simultaneously.

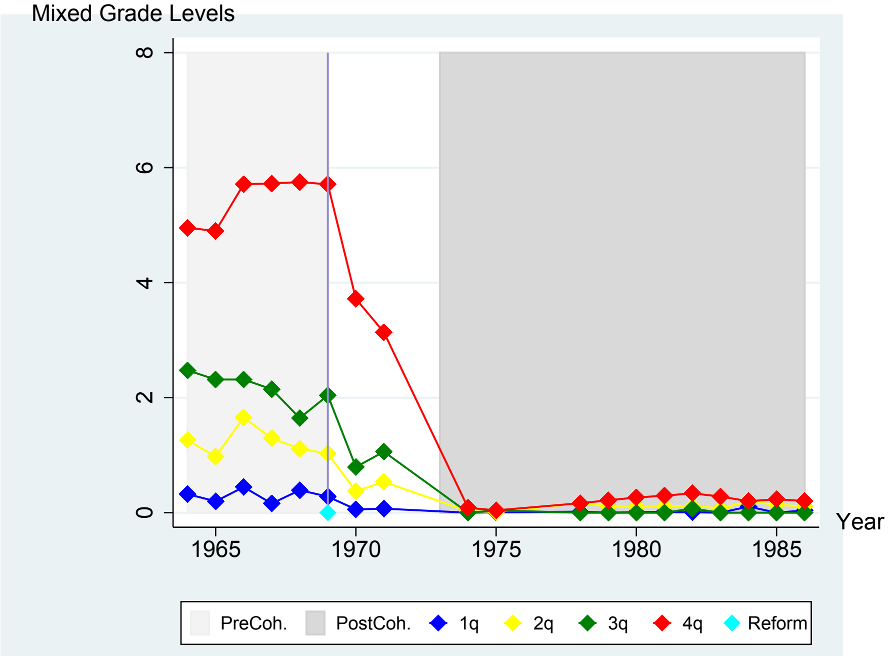
Source: Schools' Index 1964-1986. Own calculations.

ability of being taught in a multigrade school) of students was dependent on how many students of the same birth cohort had the same denomination – due to the legal obligation to teach Catholics and Protestants separately.[11] In sum, 75% of schools in the Saarland resolved to a multigrade structure prior to the reform in 1969, all of which were schools in more rural regions. Denominational schools in more urban regions, by contrast, were characterized by a single-grade structure.

The reform of 1969 had a direct impact on schools offering basic education. Inducing a change in students' distribution across school types it also indirectly affected higher education though. When denominational schools were legally abolished in various states all over Germany, this raised hot debates and interventions on behalf of the church and

[11] Verfassung des Saarlandes (1947) Art. 27 (Amtsbl. des Saarlandes, Nr. 41) Vom 05.11.1969, *available at* `http://www.verfassungen.de/de/saar/saarland47-index.htm` (23 May 2015).

Figure 1.2: Mixed Grade Levels by Treatment Probability over Time: Catholic Students



Notes: This figure shows, for the case of Catholic students, the prevalence of multigrade teaching (diyplaying the number of mixed grade levels) over time by treatment probability (in quartiles). The treatment probability depends on the number of schools in a municipality-denomination-cohort-cell that were offering multigrade teaching prior to the reform.

Source: Schools' Index 1964-1986. Own calculations.

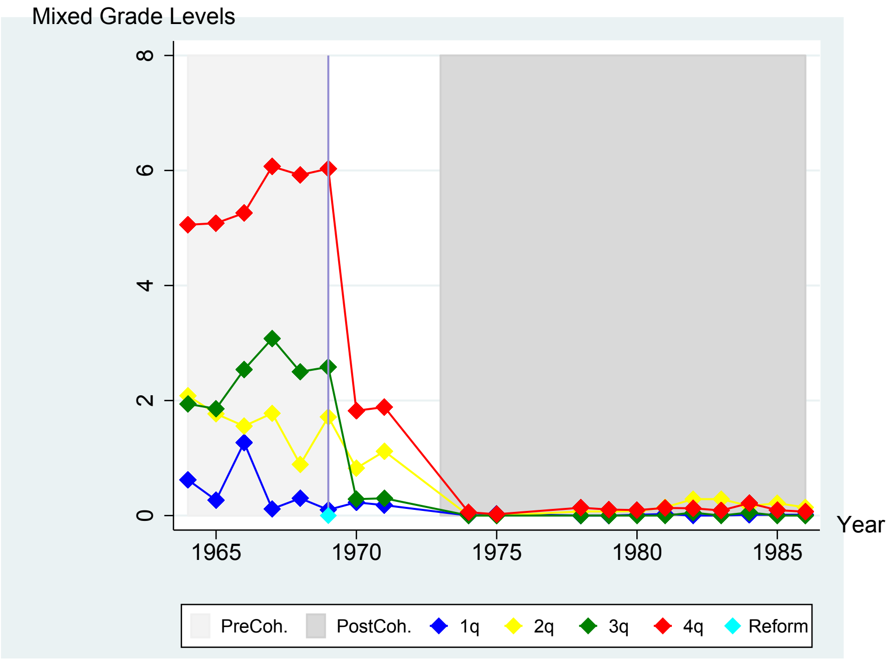
parents likewise[12] but in the Saarland the reform was carried out neatly.

Due to the reform the number of multigrade schools decreased by two thirds in less than a year and from 1974 onwards the share of multigrade schools was negligible. Thus, the reform changed the learning environment for children in more rural regions where multigrade schools predominated prior to the reform in 1969 substantially. Tiny schools were wrapped up into normal-size ones reducing the number of village schools by more than 50% while diminishing the frequency of more urban schools only moderately. In consequence, from 1974 onwards the prevalence of multigrade teaching was close to zero in both treated and control regions, see Figure 1.2 for the development of multigrade teaching in Catholic schools over time and Figure 1.3 for the case of Protestant schools.[13]

[12] `http://www.spiegel.de/spiegel/print/d-46369565.html` (01 May 2015).

[13] Tables 1.A.1, 1.A.2 and 1.A.3 in the appendix compare the number of mixed grade levels in treated and control regions prior and after 1969 separately for Catholic, Protestant and (the few) non-denominational schools.

Figure 1.3: Mixed Grade Levels by Treatment Probability over Time: Protestant Students



Notes: This figure shows, for the case of Protestant students, the prevalence of multigrade teaching (diyplaying the number of mixed grade levels) over time by treatment probability (in quartiles). The treatment probability depends on the number of schools in a municipality-denomination-cohort-cell that were offering multigrade teaching prior to the reform.

Source: Schools' Index 1964-1986. Own calculations.

The abolition of denominational schools left some villages without an own school altogether and required their children to become commuters. Having to commute anyway changed relative commuting costs to higher education schools that might previously have been prohibitive. Attending a restructured *Volksschule* or even opting for a higher education school, either way rural students were taught in much more homogeneous classes.

All key features of schools are summarized in Tables 1.1 and 1.2, partitioning the universe of *Volksschulen* into four groups, namely treated and control schools, each before and after 1969 (and separately for Catholic and Protestant schools).[14] As the tables show, by construction the reform reshaped the educational infrastructure in multiple ways and also implied more students and more teachers per school in absolute terms (EENEE, 2015). For example in the case of Protestants living in treated municipalities where multigrade

[14] The key features of non-denominational schools are shown in Table 1.A.4 in the appendix.

Table 1.1: School Characteristics by Treated and Control Status of Catholic Students

	PRE REFORM				**POST REFORM**			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Class Size	37.509	34.42	-3.089	(-9.987)	23.24	21.701	-1.539	(-8.478)
Pupils/Teacher	36.354	34.621	-1.733	(-5.678)	20.076	21.002	.926	(4.471)
Pupils/School	369.435	109.818	-259.617	(-47.896)	284.636	127.83	-156.806	(-28.302)
Girls' Share	.527	.49	-.037	(-6.04)	.48	.49	.01	(4.738)
Female Teachers' Share	.459	.427	-.033	(-3.965)	.526	.529	.004	(.471)
Teachers/School	10.125	3.056	-7.069	(-48.393)	14.292	6.155	-8.137	(-29.917)
Observations	1216	1021			2667	872		

Notes: A student is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only Catholic students and the schools they attended are considered.

Source: Schools' Index 1964-1986. Own calculations.

Table 1.2: School Characteristics by Treated and Control Status of Protestant Students

	PRE REFORM				**POST REFORM**			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Class Size	32.409	31.009	-1.4	(-3.193)	23.045	22.344	-.7	(-3.913)
Pupils/Teacher	31.465	31.404	-.061	(-.105)	20.336	20.215	-.122	(-.599)
Pupils/School	270.118	88.962	-181.156	(-27.511)	252.835	226.88	-25.955	(-4.335)
Girls' Share	.51	.494	-.016	(-1.933)	.483	.483	0	(-.06)
Female Teachers' Share	.517	.463	-.054	(-4.265)	.526	.528	.002	(.288)
Teachers/School	8.607	2.772	-5.835	(-27.904)	12.599	11.414	-1.185	(-3.986)
Observations	374	448			2607	932		

Notes: A student is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only Protestant students and the schools they attended are considered.

Source: Schools' Index 1964-1986. Own calculations.

teaching was the norm prior to the reform, average school size increased from 89 students per school to 227 students per school and from 2.8 teachers per school to 11.4 teachers per school (see Table 1.2). At first sight surprisingly, average class size shrank because the inflow of remote area children into more urban school districts was mitigated by a demographic decline in enrollment. It drastically reduced overall class size from 39 (1964) to 19 (1986) students on average, but the relative change was identical for treated and control regions. However commuting students coming from remote areas might have encountered higher quality peers from more urban municipalities (Leuven and Rønning, 2016).

For the comparison between treated municipalities (where multigrade teaching was the norm prior to 1969) and control municipalities (where single-grade teaching was the norm prior to 1969) to make sense a common trend between those regions is essential. The 1960s are called the decade of educational expansion and changes over time are indeed tremendous. We exploit that the reform eradicates multigrade classes which creates an asymmetry between otherwise parallel worlds. The following important education laws in the Saarland are all implemented well before the reform is rolled out in 1969 and they maintain a common denominator for treated and control municipalities – those with and those without a history of multigrade schools – over time.

To begin with the *Compulsory School Entry Age* fixes enrollment into primary school to age six with minor exceptions referring to each June's 30th as cut-off date.[15] Next *Compulsory Schooling Duration* requires that students stay in school for at least nine years and passing the ninth grade is rewarded with a lower secondary degree. It turns out that roughly 4:1 students finish a ninth grade already before the law inures in 1965 (Pischke and Wachter, 2005). However its implementation requires two short school years that actually compress schooling duration in 1966/67. Then, *No Tuition Fees* guarantee basic education to be free of charge, independent of the school being state- or church-maintained.[16] It limits the influence of parents' financial constraints and prevents a selection by the fee itself. Finally, *Limited School Choice* of the parents is achieved by allocating students over schools based on catchment areas.[17] To choose a certain *Volksschule* by its reputation would require the household to move into that school's catchment area. Rothstein (2006) investigates parental preferences over school choice and establishes that peer groups matter even more than schools' effectiveness. This underlines the importance of student allocation by catchment areas because it mitigates parental choice effects which interfere with the core mechanism of multigrade classes. Jointly these laws provide accuracy in comparing schooling circumstances. This is an advantage compared to class composition studies of developing countries.

We analyze a period of more than two decades of schooling conditions. Our setup is robust to symmetric shocks. When screening the most influential historical events that could have had asymmetric impacts on treated and control municipalities, a primary concern relates to fluctuations in economic activity centered in urban regions. The coal and steel crises depressed the Saarland even more than the rest of Germany (Lichtblau, 2009). They caused dramatic peaks in unemployment and overshadowed positive shocks such as the construction of the Ford plant or the infrastructure improvement by the Saar Canal.

[15] §2 Satz 1 Gesetz Nr. 826 Schulpflichtgesetz *available at* http://sl.juris.de/cgi-bin/landesrecht.py?d=http://sl.juris.de/sl/gesamt/SchulPflG_SL.htm#SchulPflG_SL_rahmen (12 June 2015).

[16] §1 Satz 1 Gesetz Nr. 662 Schulgeldfreiheit *available at* http://sl.juris.de/cgi-bin/landesrecht.py?d=http://sl.juris.de/sl/gesamt/SchulGFrhG_SL.htm (12 June 2015).

[17] §29 Satz 2 Schulordnungsgesetz vom 5. Mai 1965.

Geographic controls measuring the distance to former major smelting works, direct access to the river, etc. are one possible solution to control for these changes. It is worth mentioning that despite of these shocks the Saarland was politically nearly perfectly stable (ibid). Only the very last year of our study's time horizon is subject to a different government, therefore we expect its influence to be limited. The advantage of exploring inner-state differences becomes obvious here. By construction, many complicating aspects like tax schedules causing potential problems in Abramitzky and Lavy (2011), etc. are taken care of from the start.

1.4 Empirical Model

The key empirical question refers to the comparison of the performance of students in a multigrade environment to a single-grade environment, which is less heterogeneous in terms of birth cohorts. We tackle this question estimating a triple differences (DDD) model that exploits exogenous variation in the probability to be a multigrade student over time, region and age group.
Let Y_{1imdcy} represent individual i's outcome in municipality m with denomination d, belonging to cohort c and age group y if she attended a multigrade school and Y_{0imdcy} otherwise.

A student is defined as treated if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. If in one municipality there was one Protestant school teaching at least two grade levels jointly in all pre-reform years, then a Protestant student will be labeled as living in a multigrade municipality. This is still true if in the same municipality there exist Catholic schools which might be single-grade schools. This definition underlies the balancing tables 1.1, 1.2 and 1.4. It ensures that within a treatment-municipality-denomination-cohort cell the probability to attend a multigrade school was 100%.[18] Yet, this definition might be overly retrictive as it dismisses multigrade exposure whenever the probability was not 100%. In other words, citing the example from above, even if only in one year prior to the reform the Protestant school obtained a single-grade structure the Protestant student will be labeled as non-treated. Therefore, building on the binary defnition we employ two alternative continuous treatment indicators in our regressions.[19] Consider a municipality with two

[18] We estimate an intentention-to-treat effect. Apart from the standard assumptions for multiple differences analyis our setup requires two non-technical assumptions. First, pre-reform denomination of student and school coincide and second, the likelihood for treated and control students to start their own household follows a common trend while they are under-age. Conditional on these assumptions the probability to be treated assigned by the binary multigrade indicator is 100%.

[19] The binary treatment indicator is used in a robustness check. The results do not provide additional

Protestant schools, school A with 90 and school B with 10 students. The school-based indicator corresponds to the share of multigrade schools, the student-based indicator to the share of multigrade students of the respective municipality-denomination-cohort cell. Table 1.3 shows which indicator behaves more conservative, in the common computational scenarios.

Table 1.3: Treatment Status by Alternative Multigrade Indicators

	Multigrade Indicator		
Multigrade School?	Binary	School-based	student-based
Case I			
Both A, B	1	1	1
Case II			
School A	0	0.5	0.9
Case III			
School B	0	0.5	0.1
Case IV			
Neither A nor B	0	0	0

Note: Fictitious example considering a municipality with two Protestant schools, school A with 90 and school B with 10 students. The continuous school-based indicator corresponds to the share of multigrade schools, the continuous student-based indicator to the share of multigrade students of the respective municipality-denomination-cohort cell.

The binary indicator underlying our balancing tests is very conservative in assigning treatment status. Thus, it is most likely to reveal significant differences that potentially create non-common trends. Nevertheless, as any binary indicator, it disregards that treatment probability is gradual. Therefore it should be modeled as a continuous variable, just as we do in our preferred specifications discussed in this paper. As Table 1.3 shows the school-based indicator computes the probability to attend a multigrade school based on the number of schools per municipality-denomination-cohort cell (MDC). The student-based indicator models the probability to attend a school within a MDC cell to be proportional to the school's size, as a proxy for its capacity to take in students. Note however that the latter need not be a better indicator per se. Smaller multigrade schools were often much more extreme in collapsing grade levels than larger schools had to be. This motivates to condition on treatment intensity, something we are still working on. Of course treatment probability and treatment intensity are two different things. This is just one example to point out that apart from school size there exist multiple factors influencing the possible multigrade experience of a student. From this perspective, the school-based indicator is

insights and are available upon request.

just a neutral and thus very useful benchmark.

We estimate the *reform effect* in a regression with $Multigrade_{md} \in [0,1]$, a continuous variable measuring the likelihood of being taught in a multigrade class, the binary variable $c \in \{Pre, Post(Reform)\}$ and the binary variable $y \in \{Young, Old\}$, and a triple interaction, reflecting the DDD estimator. *Post* equals one for observations of the 1987 Census and zero for 1970. *Young* equals one for people aged fifteen to twenty in either census year and is zero for people aged 32 to 37 years.

$$\begin{aligned} Y_{imdcy} = \beta_0 + \beta_1 Multigrade_{md} + \beta_2 Post_c + \beta_3 Young_y \\ + \beta_{12} Multigrade_{md} Post_c + \beta_{13} Multigrade_{md} Young_y + \beta_{23} Post_c Young_r \\ + \beta \underbrace{Multigrade_{md} Post_c Young_y}_{D_{mdcy}} + \psi_m + \epsilon_{imdcy} \end{aligned} \tag{1.1}$$

To account for time-invariant confounders at the municipality level, we include municipality fixed effects ψ_m. To allow for correlation of errors within municipality we cluster on the municipality level (335 clusters).

Identification is thus based on the contrasts across municipalities with a different coverage of multigrade schools prior to the reform, age groups, and time. We estimate the DDD baseline reform effect including just the main effects *Multigrade, Post, Young* and their interaction terms.

We proceed by estimating the multigrade effect in more extensive specifications that include additional individual controls from population census data. These include *Age, Age Square, Young at School Entry, Female, Catholic* and *German. Young at School Entry* relates birth month and school entry cutoff date to indicate if a student is relatively young within her cohort. Combining this with administrative data from school records allows us to include additional controls. These comprise municipality-denomination-cohort level regressors *Class Size, School Size* (defined as the number of students) *Girls' Share* and *Female Teachers' Share.* We furthermore account for *Potential Commuting Costs* which we define as the average distance to the nearest *Realschule* or *Gymnasium* net of the distance to the nearest *Volksschule.*

The identifying assumption of our DDD strategy is that multigrade exposure is as good as randomly assigned conditional on observables and unobservable-but-fixed confounders. Adding a control group of elder people nets out region-specific changes that are not rooted in schooling conditions themselves. An example would be a boost in multigrade municipalities' neighborhood quality induced by state-level interventions to counteract drift to the cities (characterized by single-grade schools). The setup still requires unobservable asymmetries in teaching effectiveness and ability differences between multigrade municipalities' and single-grade municipalities' students to be time-constant, because – with only two periods in which region-specific outcomes are measured – trends are not identified, a drawback detailed in Stephens and Yang (2014). Moreover we rely on the aforementioned student allocation via catchment areas to ensure that students do not choose their school, and thus their multigrade exposure. To sum up, for multidimensional differencing to be applicable group composition needs to be spatially stable as well as groups should follow a common trend over time. Furthermore we assume *zero conditional mean, additive separability* and a *constant, weakly monotone causal effect β.*

1.5 Data

This section describes the data. Via municipality codes we combine two censuses and one schools' statistics, all of which are comprehensive, high-quality administrative datasets.[20]

Outcomes[21]

We construct schooling and labor market outcomes using individual-level census data from 1970 for the baseline and from 1987 for the follow-up cohorts. The data is available via remote execution at the German Federal Statistical Office. To evaluate final grade attainment we consider two separate dummies, namely (1) attainment of *Mittlere Reife or Fach-/ Abitur* (i.e. at least an intermediate secondary degree) and (2) attainment of *Fach-/ Abitur* (i.e. at least a high-school degree). Looking at grade attainment instead of years of schooling reflects longer schooling net of grade repetition and also identifies dropouts (EENEE, 2015). There are no test scores in the data. If there were, however their predictive power might have been limited anyway by grading on a reference curve, especially in a multigrade class, because relative grading depends on the presence of more advanced peers (Leuven and Rønning, 2016). Importantly, peer effects may trigger social competences not captured by test scores but perhaps reflected in post-schooling attainment. We therefore also use labor market outcomes to assess lasting or reemerging effects of schooling similar to Chetty et al. (2014a). In order to analyze labor market participation we use binary indicators on unemployment and labor market participation. Given labor market entry we distinguish further between blue- and white-collar occupations to capture the socio-economic status of the occupation. Note that wages are not reported in the Census 1987.[22] Table 1.4

[20] Volkszaehlungsgesetz 1970 vom 14. April 1969 (BGBl. I S. 292); Volkszaehlungsgesetz 1987 vom 8. November 1985 (BGBl. I S. 2078).

[21] Nearly all our outcomes are binary. Accordingly, the OLS regressions represent linear probability models (LPMs) which means that causality draws on the CIA, predictions may violate the [0,1] range and the error term is heteroskedastic (Angrist and Pischke, 2008).

[22] For a follow-up version of this paper, we consider to assign a standard income range based on each observation's meticulously reported profession (ISCO 88) for income mobility analysis in the sense of Chetty

shows descriptive evidence on the differences between treated and control individuals with respect to their schooling and labour market outcomes. It shows that treated individuals prior to the reform were less likely to hold at least a *Realschule* degree (RS degree) than control individuals. Furthermore, they were more likely to have a blue-collar job and less likely to have a white-collar job. According to the descriptive statistics, these differences were less pronounced after the reform. In fact, after the reform treated individuals are more likely to hold at least a *Realschule* degree than control individuals.

Table 1.4: Descriptive Statistics: Treatment, Outcomes and Controls

	PRE REFORM				**POST REFORM**			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Treatment Indicators								
MDC MG School Share	.259	1	.741	(397.086)	.028	.122	.094	(59.844)
MDC MG Pupil Share	.088	1	.912	(821.797)	.005	.064	.06	(53.115)
Outcomes								
At least RS Degree	.094	.08	-.014	(-5.298)	.371	.392	.021	(3.751)
At least A-levels	.009	.007	-.002	(-1.834)	.067	.069	.002	(.614)
Employed	.651	.653	.001	(.328)	.688	.707	.019	(3.674)
Non-Participant LM	.071	.07	-.001	(-.349)	.045	.032	-.013	(-5.694)
Blue-Collar Job	.514	.548	.034	(7.485)	.525	.538	.013	(2.313)
White-Collar Job	.407	.364	-.043	(-9.674)	.428	.428	0	(-.019)
Controls								
15-17 Year-olds	.417	.43	.013	(2.919)	.218	.227	.009	(1.966)
1 VS in MDC cell	.297	.902	.604	(156.353)	.325	.82	.495	(97.561)
Mun: max.5000 inh.	.233	.882	.649	(178.032)	.307	.893	.586	(121.127)
Female	.498	.488	-.011	(-2.376)	.449	.435	-.014	(-2.442)
Age	17.846	17.794	-.052	(-3.58)	18.566	18.518	-.048	(-3.276)
Young Within Cohort	.396	.402	.007	(1.489)	.372	.38	.008	(1.513)
Catholic	.804	.692	-.112	(-30.104)	.804	.682	-.123	(-26.083)
Protestant	.187	.292	.106	(28.829)	.17	.277	.107	(23.929)
German	.967	.979	.012	(7.898)	.952	.968	.016	(7.055)
Single	.895	.893	-.002	(-.75)	.944	.951	.007	(2.654)
Household Size	4.376	4.65	.274	(15.244)	3.742	4.039	.297	(19.378)
MDC Class Size	37.037	34.447	-2.59	(-77.175)	23.215	22.337	-.878	(-52.595)
MDC Pupils	380.272	133.094	-247.178	(-252.53)	296.094	170.926	-125.168	(-112.362)
MDC Girls Share	.531	.494	-.037	(-71.901)	.482	.486	.005	(22.21)
MDC Fem.Teachers Share	.477	.405	-.072	(-73.767)	.531	.514	-.016	(-11.977)
Commuter to VS	.045	.173	.128	(55.238)	.03	.339	.308	(96.759)
Commuting to VS (km)	.132	.521	.389	(48.743)	.054	.996	.942	(55.032)
Commuting to RS (km)	3.045	6.412	3.368	(82.812)	1.909	3.915	2.006	(53.953)
Commuting to Gym (km)	2.604	6.383	3.779	(95.454)	2.672	5.12	2.448	(50.949)
Commuter	.566	.664	.098	(21.521)	.649	.71	.062	(11.168)
Observations	54465	15694			30245	10456		

Notes: In this table, we differentiate between control and treated students (between 15 and 20 years old) pre and post to the reform in 1969. A student is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. MDC = municipality-denomination-cohort, MG = multigrade, VS = Volksschule, RS = Realschule, Gym = Gymnasium, LM = labor market.

Source: Integrated dataset of Census 1970 and 1987 and Schools' Index 1964-1986. Own calculations.

et al. (2014b).

Treatment Indicator

We determine each individual's likelihood for having been a multigrade student – considering each individual's municipality and denomination – computing two alternative continuous treatment indicators as explained in Section 1.4. The school-based indicator corresponds to the share of multigrade schools, the student-based indicator to the share of multigrade students of the respective municipality-denomination-cohort cell (MDC).[23] Table 1.4 shows that on average 26% of those students defined as control by the *binary* indicator are assigned a positive treatment probability by the *school-based* indicator. In contrast, 8.8% of those students defined as control by the *binary* indicator are assigned a positive treatment probability by the *student-based* indicator.

Controls

Using data from Saarland's Statistical Office, we obtain records on all primary and lower secondary schools from 1964 to 1986. Key figures like the numbers of male and female students and teachers, the number of classes, school's type, denomination and address are given for each school on an annual basis yielding more than 7500 school-year observations.[24] The school's address enables us to average over schooling conditions of schools of a given denomination in a given municipality in a given year. We then group the years into pre and post reform and match them to individuals in the baseline and follow up cohorts respectively via the municipality code while also considering an individual's denomination.[25] Importantly, for 80% of all schools (attended by roughly 50% of all students) a unique mapping between a student of a given denomination and the school of her denomination is possible (i.e. there is no need to match the student to an average of school characteristics of two or more schools of her denomination).

By help of the schools' records we compute pre- and post-reform municipality-denomination-cohort (MDC) averages of class size, student-teacher ratio, school size (in terms of number of students), girls' share and female teachers' share. Table 1.4 compares the main schooling characteristics between schools in treated and control municipalities. Importantly, class size, the principal rivaling input when estimating the effect of multigrade schools, is a bit lower in treated regions (on average, there were 2.6 students less per class). Since a smaller class size has presumably beneficial effects on students' achievement, this fact will

[23] See Table 1.3 for gaining an intuition of the different behavior of both indicators.

[24] We exclude special schools. Records for the years 1971/72 are missing completely. For 1966 one fifth of the data is missing but without region-specific missing patterns.

[25] In order to calculate average post-reform schooling conditions, we take schools' records from 1973-1986 into account. The cohorts of interest analyzed out of the 1987 Census are at most 20 years old in 1987 implying they entered primary school earliest in 1973.

rather lead to underestimating the effects of the abolition of multigrade classes when not controlling for class size.

The census data provide us with a set of individual-level controls all displayed in Table 1.4, most of which are commonly used and self-explanatory. The differences between treated and control individuals are in line with expectations: Treated individuals are more likely to live in municipalities with less than 5,000 inhabitants (88% vs. 23%), and are more likely to have only one Volksschule (VS) in their municipality-denomination-cohort cell (MDC), namely by 90% vs. 30%. Moreover, treated individuals are less likely to be Catholic (70% vs. 80%).

Here we briefly discuss those controls with non-standard implications. In our setting, some standard controls like household size and marital status are potentially bad control because the reform likely affects marriage and/or fertility behavior (Lundborg et al., 2012). The bad control case is even more pronounced for potential commuting costs. Students forced to commute are facing different effort costs than those attending school in direct vicinity. So continuing school at all is decided on altered premises. Simultaneously the implicit 'vicinity bonus' of lower secondary schools over higher education schools disappears in rural regions. Commuting anyway, ability-based school choice seems more natural than it has been with a *Volksschule* at walking distance and higher education schools at multiple kilometers' distance. Therefore we control for the distance to the nearest *Realschule* and/or *Gymnasium*. Importantly, however, we only include household size, marital status and commuting costs in an extended version of our regressions because we cannot rule out they are bad controls.

Sample Restrictions

Census data virtually cover all Saarlanders in each of the two survey years providing us with an unrestricted sample exceeding two million observations. We drop individuals younger than fifteen years because that is the minimum age for the outcomes we observe. Furthermore it is crucial to drop individuals between 21 and 32 years for two reasons.

First, before turning 21, people are still underage[26] such that their mobility is low. This matters because census data provide the municipality code of current residence and of school attendance. Fortunately, the residence-of-household definition ties children to their parents' address until they begin their own household.

Nevertheless, concerned with individuals moving reform-induced away from more rural regions (characterized by a higher likelihood of offering multigrade teaching) to urban

[26] Legal definition as of 1970. For a subset of outcomes we run robustness checks restricting the sample to below 18 years, the legal threshold valid in 1987. This imitates what Lundborg et al. (2012) do facing the same problem.

regions we impose that underage restriction. It leaves us with a sample of main interest consisting of five consecutive birth cohorts with individuals who are between fifteen and twenty years old in either census. All of them attend primary and lower secondary school either strictly before or strictly after the reform takes place.

Second, although there is no panel structure at the individual level, observations of the 1970 Census reappear in the survey of 1987. Individuals between 32-37 years olds in 1987 have been past schooling age already in 1970 and are therefore untreated in either census. By construction their mobility cannot change reform-induced, so it is safe to include them as a control group. However the case is much more complicated for individuals between 21 and 32 years old in 1987. They have been partially treated because they are still in lower secondary school when the reform is rolled out in 1969. With respect to multigrade exposure they fall into a transition period with exceptional schooling conditions due to fundamental restructuring. Therefore, we exclude them from our sample. Note that the seventeen-year elapse between both censuses is just short enough to preclude that parents of the post-cohorts have already been treated. Otherwise multi-generational class composition effects could accumulate, a channel established in Lundborg et al. (2012). Admittedly, the framework cannot rule out general equilibrium effects, a caveat that needs further investigation.

We furthermore restrict the sample to individuals for whom we have information on the outcomes of interest. In the end, our final dataset consists of 287,153 individuals when combining both age groups. When taking only into account the younger individuals of both censuses (aged between 15-20 years) the sample consists of 111,081 individuals.

1.6 Results

This section presents estimates of the impact of the abolition of multigrade schools on schooling and labor market outcomes. Our findings are in line with the literature suggesting a negative net effect from multigrade classes whenever other education inputs are not adapted accordingly. We show that results are robust to the inclusion of a wide range of individual characteristics and schooling covariates. Moreover we stratify the sample to investigate heterogeneity of the multigrade effect across subgroups. Throughout, we show (1) estimates of the DID estimation (i.e. not including the 32-37-year-olds as control group) using the school-based multigrade indicator, (2) estimates of the DDD estimation using the school-based multigrade indicator, (3) estimates of the DDD estimation using the student-based multigrade indicator. The multigrade indicators are calculated from the share of multigrade schools and multigrade students respectively. The latter respects the number of students (school size) upon averaging. Both indicators are continuously defined over 0 and 1 and measure each individual's multigrade exposure/treatment probability precisely.

Overall Results

Schooling Outcomes

Table 1.5 presents the main results based on the whole sample. We show estimates of the two different continuous DDD specifications (using the student-based multigrade indicator and the school-based multigrade indicator respectively) as well as estimates of a DID specification (using the school-based multigrade indicator). For each specification, we show estimates of the baseline approach (not including any controls), the core controls approach (not including potentially bad controls, see Section 1.5) and the extended controls approach (including all controls).[27] DID as well as DDD regression results displayed in Table 1.5 suggest that the abolition of denominational schools favorably influenced degree attainment. This finding is remarkably robust across our different specifications. According to the estimated coefficients the change from a multigrade school system to a single-grade school system significantly raised the average probability of attaining an intermediate secondary degree (*Mittlere Reife* or *Abitur*) by 7-11 percentage points (ppt), depending on the specification. The effect on having attained a high-school degree (*Abitur*) is also positive and indicates that the switch to a single-grade school system led to an increase of students holding a *Abitur* of around 5 ppt. A natural explanation of this finding would be that individuals spend more time on schooling because single-grade classes improve basic training. This in turn makes superior educational attainment accessible.

Professional Outcomes

The estimates in Table 1.5 show that the reform did not change the overall probability of being employed. Yet, we observe a reform-induced increase of the likelihood of holding a white-collar job and a reform-induced reduction of the likelihood of becoming a non-participant in the labor market (in other words, in the case of women, becoming a housewife). Interestingly, the labor market estimates get more precise and larger when adding the control group of elder people, i.e. turning from the DID-estimation to the DDD-estimation. This indicates that the increased take-up of white-collar jobs is not due to a region-specific labor market trend. The global gain in white-collar employment seems to be partly driven by female labor market participation which is reflected in the

[27] We present the overall results for all three approaches. In the cases of the sensitivity analysis and the subgroup analysis, however, we only display the results of regressions including the core controls. The results of the other specifications are available upon request.

Table 1.5: Overall Effects on Schooling and Labour Market Outcomes of 15-20-Year-Olds

	Schooling		**Labor Market**			
	M.Reife/Abitur	Abitur	Employed	Blue-Collar	White-Collar	Non-Participant
DDD (pupil-based)						
Baseline	0.112***	0.0375	0.00277	-0.0165	0.0473	-0.0254
	[0.0367]	[0.0242]	[0.0395]	[0.0284]	[0.0311]	[0.0194]
Core Controls	0.112***	0.0407*	0.0187	0.00427	0.0463	-0.0457**
	[0.0363]	[0.0239]	[0.0446]	[0.0253]	[0.0304]	[0.0191]
Extended Controls	0.111***	0.0409*	0.0153	-0.00264	0.0463	-0.0391**
	[0.0367]	[0.0235]	[0.0423]	[0.0250]	[0.0287]	[0.0160]
DDD (school-based)						
Baseline	0.0903***	0.0520**	0.0296	-0.0285	0.0529**	-0.0194
	[0.0290]	[0.0229]	[0.0330]	[0.0241]	[0.0267]	[0.0171]
Core Controls	0.0898***	0.0534**	0.0371	-0.0162	0.0528**	-0.0320**
	[0.0286]	[0.0225]	[0.0348]	[0.0233]	[0.0262]	[0.0162]
Extended Controls	0.0912***	0.0529**	0.0396	-0.0193	0.0514**	-0.0280**
	[0.0287]	[0.0223]	[0.0333]	[0.0227]	[0.0242]	[0.0140]
DID (school-based)						
Baseline	0.0868***	0.0117	0.0515	-0.0242	0.0543*	-0.0164
	[0.0281]	[0.0122]	[0.0384]	[0.0271]	[0.0296]	[0.0130]
Core Controls	0.0823***	0.0115	0.0443	-0.00480	0.0358	-0.0187
	[0.0286]	[0.0124]	[0.0359]	[0.0264]	[0.0303]	[0.0116]
Extended Controls	0.0817***	0.0113	0.0324	-0.00750	0.0354	-0.0178**
	[0.0281]	[0.0121]	[0.0365]	[0.0261]	[0.0278]	[0.00794]
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
N (DDDpupil)	287153	287153	287153	287153	287153	287151
N (DDDschool)	287153	287153	287153	287153	287153	287151
N (DID)	111081	111081	111081	111081	111081	111079
Cluster (DDDpupil)	337	337	337	337	337	337
Cluster (DDDschool)	337	337	337	337	337	337
Cluster (DID)	333	333	333	333	333	333
Adj.R2 (DDDpupil)	0.129	0.0797	0.276	0.289	0.0931	0.510
Adj.R2 (DDDschool)	0.129	0.0797	0.276	0.289	0.0931	0.510
Adj.R2 (DID)	0.181	0.0660	0.172	0.234	0.189	0.550

Notes: This table shows in the upper part estimates of the DDD estimation using the student-based multigrade indicator, then it shows the estimates of the DDD estimation using the school-based multigrade indicator and in the bottom part it shows the estimates of the DID estimation (i.e. not including the 32-37-year-olds as control group) using the school-based multigrade indicator. The multigrade indicators are calculated from the share of multigrade schools and multigrade students respectively. For each specification, the table shows estimates of the baseline approach (not including any controls), estimates of the core controls approach (not including potentially bad controls) and estimates of the extended controls approach (including all controls).

Standard errors are clustered at the municipality level and are shown in parenthesis: *p < 0.10, ** p<0.05, ***p<0.010.

Source: Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986. Own calculations.

housewife/non-labor-market-participation status declining by 3 ppt. Below, we discuss channels of gender-specific responsiveness to the treatment in more detail. In sum, results suggest that reform-induced higher educational attainment led to an increase of better qualified employment.[28]

[28] The importance to assess general equilibrium effects for policy recommendations is detailed in Heckman et al. (1998). As mentioned before the sizable period elapsing between pre- and post cohorts' outcomes heightens the probability that general equilibrium effects understate or overstate positive effects from

Sensitivity Analysis

Table 1.A.5 in the appendix shows the results of the main regressions – using the core controls approach – when restricting the sample in two different ways. For the sample used for regressions in the upper part of Table 1.A.5, we only take into account individuals for whom the municipality where they went to school is definitely known, i.e. we can exclude migration in order to take up employment elsewhere. This implies that this group of individuals represents a negative selection – they might be more afraid to move away from home or do not have sufficiently good skills to get employed elsewhere. The results in Table 1.A.5 are in line with this negative selection argument. While we observe a similar reaction to the switch from multigrade to single-grade teaching in terms of the attainment of a higher secondary degree, the labor market response is much smaller than for the whole sample. For results in the bottom part of Table 1.A.5, we restrict the sample to those individuals who live in those municipalities in which a unique mapping between individual and school is possible (since there is at maximum one school per denomination prior to the reform). This restriction makes a clean attribution of school controls possible. The disadvantage of this restriction is that we are left with the very small municipalities, and face, again, the problem of negative selection: those students staying in small villages are probably less ambitious. The results in Table 1.A.5 are very similar to the overall findings in Table 1.5. In contrast to the upper part of Table 1.A.5 we also find a significant negative effect on the likelihood to become a housewife/non-participant in the labor market.

Subgroup Analysis

Related studies motivate robustness checks by gender and denomination which we present in the following.

Boys & Girls

While the reasons for gender-specific reactions to education policies are still debated their existence has been shown repeatedly. Along these lines Angrist and Lavy (1999) find incentives pushing college certification rates only for Israeli girls. Deming et al. (2014) document gender-dependent attainment gains in US post-secondary education where only girls respond to higher school quality. These findings are complemented by relatively higher female responsiveness to tracking (Duflo et al., 2011). However Whitmore (2005) draws on the STAR experiment to single out gender-neutral gains by class size reduction. As shown in Table 1.A.6 in the appendix, Saarland's data confirm girls' final grade attainment to

improved education. Disentangling the partial effect we are interested in and the general effect offsetting it requires a joint estimation of skill supply and demand elasticity. The latter lies - for now - beyond the scope of our study.

improve more strongly than that of boys in the case of secondary education. While the switch from a multigrade system to a single-grade system led to a 11-16 ppt increase in a girl's likelihood to attain at least a secondary degree, it increased a boy's likelihood to attain such a degree by only 5-8 ppt which is already strong. Regarding the probability of attaining at least a high-school degree (*Abitur*), however, girls fare somewhat worse. Interestingly, as regards labor market outcomes, we do not observe large differences across gender and, moreover, the coefficients are not significant when splitting the sample. Yet, results in Table 1.A.6 show that the switch from a multigrade school system to a single-grade school system decreased the likelihood of becoming a housewife/non-participant in the labor market significantly for girls, but not for boys. What are potential explanations for girls benefiting more than boys from the disappearance of multigrade teaching? One possibility refers to girls being on average higher achieving than boys. Analogously it could be that their trajectories of improved education inputs are steeper. The literature also suggests girls to be less competitive than boys (Leuven and Rønning, 2016). Thus learning in highly heterogeneous multigrade groups might be more demanding for them. Consequently, they profit more from the switch to single-grade classes.

Catholics & Protestants

Table 1.A.7 in the appendix shows the estimated coefficients for the sample stratified by denomination. Overall, it indicates that both groups of individuals benefited from the reform in terms of their educational outcomes. Surprisingly, Protestants seem to have gained by much more than Catholics did. Moreover, Table 1.A.7 shows insignificant and close-to-zero labor market effects for Catholics, while it indicates large and significant reform-induced gains for Protestants. What are potential explanations of this finding? Again, as in the case of explaining larger benefits of the reform accruing to girls than to boys, it could be that Protestant students are on average higher achieving than Catholic students and are therefore responding more to an increase of inputs into their education production function. This touches upon the Weber Hypothesis of Protestants' inherently superior work ethics, see Becker and Woessmann (2009) who connect wide-spread literacy to Protestants' prosperity. In a follow-up version of this paper, we will offer more evidence to gain a deeper understanding of the reasons for the heterogeneity of our findings with respect to denomination.

1.7 Conclusion and Outlook

This paper addresses the question how attending a multigrade school affects school attainment and labor market outcomes, and whether there are any differences by gender or denomination in this effect. To answer this question our analysis exploits the abolition of Saarland's denominational schools as a natural experiment that overcomes the main challenges of impact evaluations for policy design (McEwan, 2008).

The reform produces a sharp treatment effect, in terms of the variation of the reduced probability to attend a multigrade class caused by an exogenous event, namely the abolition of denominational schools. Based on a legal change that is rapidly and comprehensively accomplished the setup provokes, if any, negligible anticipation or conditional-on-participation effects. Highly accurate school-level data allow us to control for rivaling changes in the educational infrastructure that are also implied by abandoning denominational tracking. The estimation approach based on triple differences plausibly identifies causal links between treatment and outcome candidates. Our results are remarkably robust across specifications and unambiguously suggest single-grade classes to be more beneficial for students' educational and labor market outcomes. Due to the reform treated students shift away from obtaining only a lower secondary degree (*Volksschulabschluss*) and a blue-collar job. Their probability to attain at least an intermediate secondary degree (*Realschulabschluss*) and to become a white-collar employee increases significantly when switching from a multigrade school system to a single-grade school system. Stratifying the main sample the emerging patterns line up with asymmetric treatment responses observed in related studies. Splitting the sample by denomination suggests that Protestant students profited more from the reform than Catholic students did. Moreover, we show that girls were more affected by the switch from a multigrade to a single-grade school system than boys. Our research approach provides external validity for the European context, which is particularly relevant in the light of the ongoing demographic change. To our knowledge this is the first study to exploit a large-scale experiment on multigrade classes in Germany. Policy interest in combination classes spans the globe but major empirical research is located in developing countries. Therefore, it suffers from limited external validity for the Eurpean context as third-world schooling bears many peculiarities. Saarland's data date back to the 1960s but the insights provided seem still easier adaptable for use in Europe. The village schools we observe are much more likely to produce positive peer effects than schools in developing countries doomed by overage-for-grade students. Our findings nevertheless suggest that a beneficial multigrade system needs strategic adjustments. We conclude that peer effects based on student collaboration alone are no panacea which refutes the argument that reallocation is a *costless* way to improve education.

Still, there are some open questions that we want to address in a follow-up version of this paper: Why do we observe stronger effects of the reform for Protestants? So far, we did not consider the pure effect of the abolition of *denominational* schools, but assume that the effects we find are the result of the disappearance of multigrade schools due to the abolition of denominational teaching. Yet, it might be that part of the *multigrade* effect is due to *denominational* teaching methods (that had a different impact in treated and control groups). Future research will thus try to disentangle the denominational effect from the multigrade effect. Furthermore, we will investigate in more depth why the shift from multigrade teaching to single-grade teaching has larger effects for girls. Using German data of the PIRLS study (Progress in International Reading Literacy Study) we will investigate whether gender-specific effects of multigrade teaching already arise at a young age. In particular, we will use the variation in the introduction of multigrade teaching in primary schools across German states between 2000 and 2010.

1.A Appendix

Table 1.A.1: Mixed Grade Levels by Treated and Control Status of Catholic Students

	PRE REFORM				**POST REFORM**			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Mixed Levels/School	.986	5.571	4.585	(57.118)	.049	.226	.177	(10.436)
Not Mixing	.704	0	-.704	(-49.25)	.977	.834	-.143	(-16.189)
Mixing Two Levels	.1	.032	-.067	(-6.304)	.012	.107	.095	(13.586)
Mixing Three Levels	.048	.045	-.003	(-.296)	.006	.06	.053	(10.011)
Mixing Four Levels	.02	.072	.053	(6.119)	0	0	0	(-.572)
Mixing Five Levels	.027	.105	.078	(7.648)	.001	0	-.001	(-.991)
Mixing Six Levels	.03	.139	.109	(9.623)	.003	0	-.003	(-1.718)
Mixing Seven Levels	.038	.245	.207	(15.104)	0	0	0	(-.572)
Mixing Eight Levels	.025	.244	.218	(16.455)	0	0	0	(.)
Mixing All Levels	.008	.118	.109	(11.312)	0	0	0	(.)
Observations	1216	1021			2667	872		

Notes: A student is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only Catholic students and the schools they attended are considered.
Source: Schools' Index 1964-1986. Own calculations.

Table 1.A.2: Mixed Grade Levels by Treated and Control Status of Protestant Students

	PRE REFORM				**POST REFORM**			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Mixed Levels/School	1.61	5.806	4.196	(29.635)	.087	.109	.023	(1.352)
Not Mixing	.58	0	-.58	(-24.854)	.94	.945	.005	(.571)
Mixing Two Levels	.078	.027	-.051	(-3.347)	.035	.034	-.001	(-.136)
Mixing Three Levels	.059	.038	-.021	(-1.402)	.024	.008	-.016	(-3.087)
Mixing Four Levels	.035	.056	.021	(1.431)	0	.001	.001	(1.673)
Mixing Five Levels	.08	.083	.002	(.124)	0	.003	.003	(2.901)
Mixing Six Levels	.067	.138	.072	(3.339)	.001	.008	.007	(3.513)
Mixing Seven Levels	.07	.252	.183	(7.165)	0	.001	.001	(1.673)
Mixing Eight Levels	.019	.25	.231	(9.919)	0	0	0	(.)
Mixing All Levels	.013	.156	.143	(7.302)	0	0	0	(.)
Observations	374	448			2607	932		

Notes: A student is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only Protestant students and the schools they attended are considered.
Source: Schools' Index 1964-1986. Own calculations.

Table 1.A.3: Mixed Grade Levels by Treated and Control Status of Non-Denominational Students

	PRE REFORM				**POST REFORM**			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Mixed Levels/School	1.29	4.259	2.969	(8.423)	.071	.239	.168	(7.613)
Not Mixing	.623	0	-.623	(-9.717)	.95	.881	-.07	(-5.943)
Mixing Two Levels	.058	.121	.063	(1.25)	.031	.064	.033	(3.611)
Mixing Three Levels	.101	.121	.019	(.342)	.018	.031	.013	(1.89)
Mixing Four Levels	.043	.103	.06	(1.31)	0	.002	.002	(2.616)
Mixing Five Levels	.029	.155	.126	(2.563)	0	.007	.007	(4.54)
Mixing Six Levels	.101	.155	.054	(.905)	.001	.013	.012	(4.865)
Mixing Seven Levels	.029	.224	.195	(3.532)	0	.002	.002	(2.616)
Mixing Eight Levels	.014	.121	.106	(2.494)	0	0	0	(.)
Mixing All Levels	0	0	0	(.)	0	0	0	(.)
Observations	69	58			3087	452		

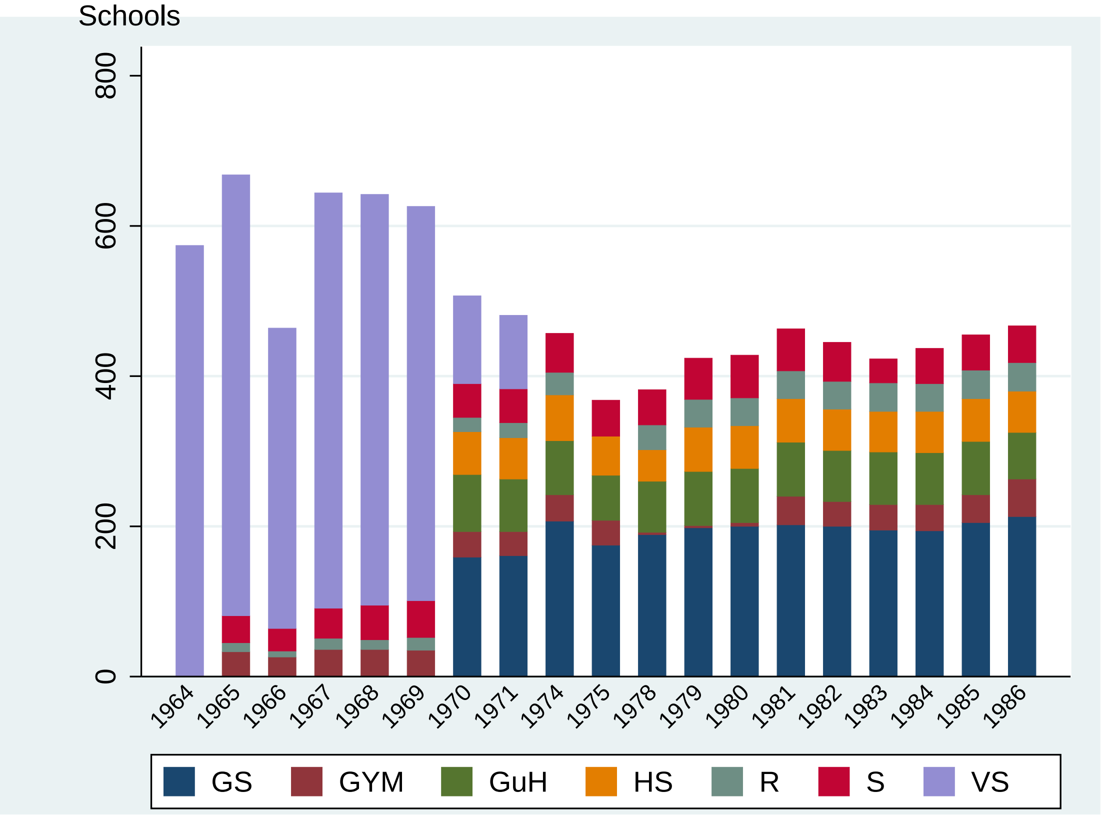
Notes: A student is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only non-denominational students and the schools they attended are considered.
Source: Schools' Index 1964-1986. Own calculations.

Table 1.A.4: School Characteristics by Treated and Control Status of Non-Denominational Students

	PRE REFORM				**POST REFORM**			
	Control	Treated	Diff.	t-stat	Control	Treated	Diff.	t-stat
Class Size	31.087	31.397	.31	(.478)	23.017	21.788	-1.23	(-5.216)
Pupils/Teacher	31.043	32	.957	(1.377)	20.312	20.252	-.06	(-.223)
Pupils/School	321.826	149.81	-172.016	(-7.972)	251.168	210.701	-40.467	(-5.127)
Girls' Share	.493	.473	-.02	(-2.031)	.483	.483	0	(.052)
Female Teachers' Share	.574	.437	-.137	(-5.42)	.522	.559	.037	(3.807)
Teachers/School	10.188	4.655	-5.533	(-8.269)	12.539	10.564	-1.975	(-5.042)
Observations	69	58			3087	452		

Notes: A student is defined as *treated* if she is living in a municipality where all schools of her denomination were multigrade schools throughout all years prior to the reform in 1969. In this table, only non-denominational students and the schools they attended are considered.
Source: Schools' Index 1964-1986. Own calculations.

Figure 1.A.1: Main School Types' Distribution Over Time



Note: Schools' Index 1964-1986 (Own calculations). Records on 1972/73 and 1976/77 are missing completely. In 1964 only the type *Volksschule* (VS) is reported. 1966 about 20% of all types are missing. 1975 there are no records for Realschule (R) and 1978-80 for Gymnasiun (GYM). GS=Grundschule, GuH=Grund- und Hauptschule, HS=Hauptschule, S=Sonderschule.

Table 1.A.5: Effects on Schooling and Labour Market Outcomes – Alternative Sample Restrictions

	Schooling		**Labor Market**			
	M.Reife/Abitur	Abitur	Employed	Blue-Collar	White-Collar	Non-Participant
CERTAIN RESIDENCE						
DDD (pupil-based)						
Core Controls	0.106***	0.0284	-0.00849	-0.00316	0.0417	-0.0308
	[0.0342]	[0.0199]	[0.0535]	[0.0337]	[0.0482]	[0.0390]
DDD (school-based)						
Core Controls	0.0660**	0.0284	0.00131	0.0169	0.00293	-0.0153
	[0.0285]	[0.0183]	[0.0422]	[0.0235]	[0.0356]	[0.0293]
DID (school-based)						
Core Controls	0.0940***	0.00913	0.0589**	0.00972	-0.00211	-0.00702
	[0.0289]	[0.0116]	[0.0258]	[0.0306]	[0.0359]	[0.0190]
UNIQUE MAPPING						
DDD (pupil-based)						
Core Controls	0.0840**	0.0251	-0.000134	-0.00807	0.0548*	-0.0439**
	[0.0352]	[0.0221]	[0.0457]	[0.0263]	[0.0314]	[0.0201]
DDD (school-based)						
Core Controls	0.0529*	0.0210	0.0114	-0.00196	0.0413	-0.0355**
	[0.0276]	[0.0184]	[0.0371]	[0.0220]	[0.0258]	[0.0162]
DID (school-based)						
Core Controls	0.0652**	0.0136	0.0155	-0.00340	0.0272	-0.0161
	[0.0259]	[0.0118]	[0.0355]	[0.0280]	[0.0311]	[0.0109]
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
CERTAIN RESIDENCE						
N (DDDpupil)	132717	132717	132717	132717	132717	132716
N (DDDschool)	132717	132717	132717	132717	132717	132716
N (DID)	62445	62445	62445	62445	62445	62444
UNIQUE MAPPING						
N (DDDpupil)	125976	125976	125976	125976	125976	125975
N (DDDschool)	125976	125976	125976	125976	125976	125975
N (DID)	48836	48836	48836	48836	48836	48835

Notes: This table shows the results when restricting the sample in two different ways. In the upper part, only those individuals are taken into account for whom the municipality where they went to school is definitely known, i.e. we can exclude migration in order to take up employment elsewhere. In the bottom part, the sample is restricted to those individuals who live in those municipalities in which a unique mapping between individual and school is possible (since there is at maximum one school per denomination prior to the reform). In each part, first estimates of the DDD estimation using the student-based multigrade indicator are shown, then the estimates of the DDD estimation using the school-based multigrade indicator and then the estimates of the DID estimation (i.e. not including the 32-37-year-olds as control group) using the school-based multigrade indicator. The multigrade indicators are calculated from the share of multigrade schools and multigrade students respectively. For each specification, the estimates of the core controls approach (not including potentially bad controls) are shown.

Standard errors are clustered at the municipality level and are shown in parenthesis: *p < 0.10, ** p<0.05, ***p<0.010.

Source: Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986. Own calculations.

Table 1.A.6: Effects on Schooling and Labour Market Outcomes – Stratified by Gender

	Schooling		**Labor Market**			
	M.Reife/Abitur	Abitur	Employed	Blue-Collar	White-Collar	Non-Participant
BOYS						
DDD (pupil-based)						
Core Controls	0.0712**	0.0467	0.0236	-0.0286	0.0248	0.00828*
	[0.0355]	[0.0285]	[0.0465]	[0.0346]	[0.0328]	[0.00442]
DDD (school-based)						
Core Controls	0.0693**	0.0681**	0.0404	-0.0494	0.0486	0.00624
	[0.0318]	[0.0277]	[0.0361]	[0.0371]	[0.0358]	[0.00433]
DID (school-based)						
Core Controls	0.0554*	0.0135	0.0671*	0.00760	0.00457	-0.00439**
	[0.0326]	[0.0160]	[0.0407]	[0.0346]	[0.0353]	[0.00191]
GIRLS						
DDD (pupil-based)						
Core Controls	0.159***	0.0384	0.00721	0.0365	0.0655	-0.0953***
	[0.0469]	[0.0243]	[0.0689]	[0.0390]	[0.0534]	[0.0343]
DDD (school-based)						
Core Controls	0.116***	0.0411*	0.0304	0.0172	0.0542	-0.0666**
	[0.0390]	[0.0241]	[0.0500]	[0.0269]	[0.0422]	[0.0325]
DID (school-based)						
Core Controls	0.116***	0.00641	0.0179	-0.0202	0.0762*	-0.0378
	[0.0350]	[0.0144]	[0.0511]	[0.0369]	[0.0452]	[0.0232]
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
BOYS						
N (DDDpupil)	146633	146633	146633	146633	146633	146631
N (DDDschool)	146633	146633	146633	146633	146633	146631
N (DID)	58042	58042	58042	58042	58042	58040
GIRLS						
N (DDDpupil)	140520	140520	140520	140520	140520	140520
N (DDDschool)	140520	140520	140520	140520	140520	140520
N (DID)	53039	53039	53039	53039	53039	53039

Notes: This table shows the results when stratifying the sample by gender. For each subgroup, first estimates of the DDD estimation using the student-based multigrade indicator are shown, then the estimates of the DDD estimation using the school-based multigrade indicator and then the estimates of the DID estimation (i.e. not including the 32-37-year-olds as control group) using the school-based multigrade indicator. The multigrade indicators are calculated from the share of multigrade schools and multigrade students respectively. For each specification, the estimates of the core controls approach (not including potentially bad controls) are shown.

Standard errors are clustered at the municipality level and are shown in parenthesis: *p < 0.10, ** p<0.05, ***p<0.010.

Source: Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986. Own calculations.

Table 1.A.7: Effects on Schooling and Labour Market Outcomes – Stratified by Denomination

	Schooling		**Labor Market**			
	M.Reife/Abitur	Abitur	Employed	Blue-Collar	White-Collar	Non-Participant
CATHOLICS						
DDD (pupil-based)						
Core Controls	0.0726**	0.0239	0.0132	0.0158	0.0120	-0.0292
	[0.0360]	[0.0244]	[0.0424]	[0.0277]	[0.0335]	[0.0199]
DDD (school-based)						
Core Controls	0.0658**	0.0354*	0.0328	-0.0000945	0.0192	-0.0173
	[0.0274]	[0.0202]	[0.0330]	[0.0238]	[0.0273]	[0.0164]
DID (school-based)						
Core Controls	0.0800***	0.00828	0.0430	-0.00836	0.0352	-0.0142
	[0.0282]	[0.0126]	[0.0358]	[0.0271]	[0.0304]	[0.0104]
PROTESTANTS						
DDD (pupil-based)						
Core Controls	0.299***	0.0335	-0.216	-0.0993	0.150	0.0131
	[0.0888]	[0.0543]	[0.174]	[0.0818]	[0.103]	[0.0664]
DDD (school-based)						
Core Controls	0.110*	0.0735	-0.0744	-0.108*	0.138**	0.00421
	[0.0588]	[0.0553]	[0.0985]	[0.0617]	[0.0609]	[0.0371]
DID (school-based)						
Core Controls	0.0332	0.0328	0.138	0.0658	-0.0303	-0.0366
	[0.0836]	[0.0386]	[0.114]	[0.0684]	[0.0796]	[0.0390]
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
CATHOLICS						
N (DDDpupil)	217373	217373	217373	217373	217373	217371
N (DDDschool)	217373	217373	217373	217373	217373	217371
N (DID)	86288	86288	86288	86288	86288	86286
PROTESTANTS						
N (DDDpupil)	61070	61070	61070	61070	61070	61070
N (DDDschool)	61070	61070	61070	61070	61070	61070
N (DID)	22837	22837	22837	22837	22837	22837

Notes: This table shows the results when stratifying the sample by denomination. For each subgroup, first estimates of the DDD estimation using the student-based multigrade indicator are shown, then the estimates of the DDD estimation using the school-based multigrade indicator and then the estimates of the DID estimation (i.e. not including the 32-37-year-olds as control group) using the school-based multigrade indicator. The multigrade indicators are calculated from the share of multigrade schools and multigrade students respectively. For each specification, the estimates of the core controls approach (not including potentially bad controls) are shown.

Standard errors are clustered at the municipality level and are shown in parenthesis: *p < 0.10, ** p<0.05, ***p<0.010.

Source: Integrated dataset of Census 1970 & 1987 and Schools' Index 1964-1986. Own calculations.

Chapter 2

Effects of Multigrade Classes in Primary Schools on Educational Outcomes

2.1 Introduction

multigrade teaching with more than one age group attending the same class is common practice in most countries around the world. A considerable proportion of pupils in primary schools in developing countries including India, Peru, Sri Lanka and Pakistan, but also in developed countries including Finland, the Netherlands, the UK and Germany, experience teaching of more than one age cohort of pupils in one classroom (see, e.g., Little, 2004; Mulkeen and Higgings, 2009, for an overview). The determinants of multigrade teaching are diverse, and range from lack of teachers, rural depopulation, and adjusting to enrollment fluctuations in the context of demographic change, to pedagogical arguments related to peer effects. While multigrade teaching has been advocated since the 1920s as a way to overcome disadvantages of single-class teaching and to foster the potential of pupil interactions, the evidence on the effects of multigrade teaching on academic performance, particularly among primary school children, is mixed. Findings of potentially detrimental effects of attending multigrade classes on subsequent outcomes has led to heated debates regarding the appropriateness of multigrade teaching as a legitimate goal of education policies (see Carle and Metzen, 2014, for a recent survey of the pedagocial literature).

This paper provides novel evidence on the causal effect of multigrade teaching in primary schools on literacy skills by the end of primary school. The analysis is based on student test score data of more than 68'000 fourth-graders from Germany. To measure educational outcomes, the analysis considers the performance of fourth-graders, which constitute an important determinant for sorting into the different secondary school tracks,

This chapter is based on joint work with Uwe Sunde and Larissa Zierow.

which typically occurs after fourth grade. We combine data originally collected within the PIRLS framework (Progress in International Reading Literacy Study) and the *IQB Laendervergleich* (a German National Student Assessment). In particular, we use test scores for reading skills at the end of fourth grade, German grades at the end of fourth grade, teachers' recommendations for the secondary school track, as well as enjoyment of school as outcomes variables. The analysis makes use of the 2001, 2006, 2011 and 2016 cohorts of fourth-graders, for whom these outcomes are observed, and matches these students with self-collected information about the respective state reforms introducing multigrade classes in primary school to obtain information about the treatment status.

The identification is based on the repeated comparison of fourth grade student cohorts from schools spread over all states of Germany. The identifying variation is the result of a natural experiment that occurred in the context of the staggered introduction of flexible school entrance levels across German states between 1997 and 2010. This experiment delivers quasi-random variation in the exposure to multigrade teaching that rules out typical concerns related to selection. The staggered introduction provides variation in treatment exposure that allows us to eliminate state and time fixed effects. Maintaining the standard common trend assumption across states the regional variation in the treatment reveals the causal impact of the experience of multigrade teaching along the lines of an intention-to-treat analysis.

The results document that exposure to multigrade teaching has detrimental effects on educational outcomes measured at age 10 (end of fourth grade). On average, multigrade teaching in the first years of primary school entails a significant and robust negative effect on reading test scores of about 6% of a standard deviation, and a significant negative effect on German grades by 1/9 of a standard deviation. The effects are more pronounced for girls and show little heterogeneity with respect to parental background characteristics.

The results of our study contribute to the literature in several ways. Early work on the effects of multigrade teaching often fails to identify causal effects because of selection into multigrade classes (Veenman, 1995; Mason and Burns, 1996).[1] To address this issue, Sims (2010) made use of an instrumental variable strategy based on class size caps imposed by the California Class Size Reduction Program and shows that multigrade classes negatively affect test scores in Grades 2 and 3. Relying on survey data and comparing non-random but observationally equivalent single-grade and mixed-age classes in Sweden, Lindström and Lindahl (2011) report a sizable negative impact. Recent work by Leuven and Rønning (2016) has made use of discontinuous grade mixing rules in Norwegian junior high schools (grade 7-9). Their results document positive effects of multigrade teaching on young students, but negative effects on more mature students within a class. Using a minimum class

[1] For a comprehensive overview of the literature on multigrade classes, see also Gerhardts et al. (2021a).

size rule in Italy which leads to multigrade classes, Checchi and De Paola (2018a) find negative effects of multigrade teaching on numeracy of fifth-graders. Our results add to this small number of studies that report causal evidence on the impact of multigrade classes by using the setting of staggered German state reforms to identify the causal impact of multigrade classes on performance of fourth-graders in Germany. In light of the ongoing debate among German education scientists, this evidence sheds new light on the effects in various dimensions.

Our evidence on the short-run effects of multigrade teaching in Germany complements the findings of a companion study on the long-term effects of multigrade classes (Gerhardts et al., 2021a). In Gerhardts et al. (2021a), we find that the abolition of denominational schools implied the abolishment of multigrade classes in the German state *Saarland* in 1969. Using this setting, we show that multigrade teaching has a causal negative impact on the students' educational and labour market outcomes measured in adult age, which is especially pronounced for women. The results presented here are consistent with these finding of negative effects of multigrade classes lasting into adulthood and document that the negative effects can be traced to the exposure to multigrade teaching during primary school.

The remaining part of the paper is structured as follows. Section 2.2 describes the institutional background. Section 2.3 provides details on the two data sources we use and presents our identification strategy. Section 2.4 presents the main empirical results and shows the results of the subgroup analysis. Section 2.5 discusses the results of several robustness tests. Section 2.6 concludes.

2.2 Institutional Background

2.2.1 The German School System

In Germany, education policy is the responsibility of federal states. This implies that each of the country's 16 federal states is solely responsible for its respective school system. Although differences exist across states, the general structure is still rather uniform. Before school, the large majority of children attends kindergarten. While only about 35% of children aged 1–3 years receive day care, 92.5% of children aged 3–6 attend kindergarten or receive another form of day care.[2] Usually at the age of six, children are enrolled in primary school. After four years at primary school, i.e., typically at age 10, the school system tracks children into three secondary school tracks: lower secondary school (Hauptschule),

[2] https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Soziales/Kindertagesbetreuung/_inhalt.html

intermediate secondary school (Realschule), and high track grammar school (Gymnasium) in which students attain the university entrance qualification (Abitur).[3] The selection into a particular track is based on ability. Teachers in primary school recommend the highest school track they deem to be suitable for the child.[4] In light of this, our analysis makes use of fourth-graders' test scores as well as of teacher recommendations for the track in secondary schools as outcomes for the analysis of the effects of multigrade teaching on educational outcomes and opportunities.

2.2.2 The Reforms under Study

Reasons for the Introduction

In 1997, the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (*Kultusministerkonferenz*) discussed several national and international studies which showed that children with low socio-economic backgrounds were disadvantaged in the German school system (Wagener, 2014). Furthermore, since the 1990s an increase in the heterogeneity of abilities and skills at the time of enrollment was observed. As a result, among others, for 8–12% of children at school entrance age, enrollment to primary school was postponed by one year. To counteract on this development, the ministers of education agreed on lowering the school entry age and introducing the so-called flexible school entry stage. This concept would have implied that children could enroll in primary school also without a school entry examination certifying their school readiness[5] and that their time in grade 1 and grade 2 could be set individually. Children of both grades would be mixed and some of them would stay in this stage for one year and others for up to three years. The pedagogical arguments for such a change included that the flexible entry stage allowed to provide special support to children at the beginning of their academic education. The hope was that multigrade classes could take into account the substantial heterogeneity in abilities, social background, and interest of enrolling children. This was supposed to happen through the optimal support of both high and low achieving students by giving the former the possibility to keep pace with students

[3] In the city state Berlin and in Brandenburg, primary school lasts for six years before selection into higher tracks.

[4] In some federal states, this recommendation acts as a limit for the schooling available to the child. Parents are subsequently responsible for choosing their child's secondary school track from the (limited) set of available school tracks.

[5] The school entrance examination is a mandatory medical screening meant to promote children's health by diagnosing medical anomalies and providing necessary treatment as early as possible. It is conducted by pediatricians employed by the local health service who document children's development including their "readiness" for primary school. The examination takes place in the year prior to entering primary school when children are on average six years old.

in second grade and hereby foster their intrinsic motivation and supporting the latter via a more intense education as well as giving them more room for their personal development. The goal of the reform was thus to smooth out potential differences in knowledge and skills directly at the beginning of the school careers in order to ensure a good knowledge in basic skills for all children.

However, in the end not all states introduced this flexible entry stage as a mandatory system. Almost all states implemented pilot projects to test the idea of having a multigrade classroom at the beginning of primary school. In sum, the flexible school entrance stage was present in approximately 20 percent of all primary schools nation-wide since the first reforms.

Reactions to the Reforms

Politicians, teachers, and parents reacted all very differently to the introduction of the flexible school entry stage. On the positive side, some hoped that age heterogeneous learning groups could especially help pupils with lower skills or knowledge, because older classmates explain topics more intuitively and less abstract than teachers. It was argued that also more advanced children could benefit from these learning groups. Furthermore, the lower level of competition among pupils due to different tasks and lacking comparison of their grades was regarded as a beneficial development.

On the negative side, however, the flexible school entry age implied more effort and preparation time for teachers. It turned out to be more difficult to teach a class with a more heterogeneous age and skill structure. Often, older pupils cannot or do not want to help their younger classmates because of lacking empathy or patience (Heinzel and Koch, 2017).

To the best of our knowledge, no causal evaluations of the flexible school entrance stage in Germany have been conducted. The existing descriptive studies do not provide a clear result on whether multigrade classrooms at the beginning of primary school have any beneficial effects and whether they reduce educational inequality (Helbig and Nikolai, 2015).

Reform Timing

After the *Kultusministerkonferenz* in 1997, all states except of one (the *Saarland*) introduced multigrade teaching in a flexible school entrance system in some pilot schools. Yet, as Table 2.1 shows, only few states introduced a flexible school entrance stage as a mandatory system.

Table 2.1: Multigrade Teaching in Flexible School Entrance Levels - Reform Overview

State	Year	Flexible School Entrance Reform	Mandatory	Optional
Baden-Wurttemberg	Since 1997	Model Projects in 82 schools	none	none
Bavaria	2010-2014 2017 2019	Model Projects in 20 schools Number of schools gradually extended until 2017 216 schools	none	none
Berlin	Since 1992 2005 2010	Model Projects in 340 schools State-wide implementation Choice between FSE and Traditional System	cohort 2011	cohorts 2011,2016
Brandenburg	1992-1995 1999-2002 2000-2004 2003 2010	Pilot Projects in 2 schools Model Projects in 2 schools Extension to 20 schools Extension to 139 schools State-wide implementation	cohort 2016	cohort 2016
Bremen	1993-1995 2005	Model Projects in 2 schools Optional for all primary schools	none	cohorts 2011,2016
Hamburg	1994-1996	Model Projects in 2 schools	none	none
Hesse	1994-1998 1998-2004 2007	Model Projects in 6 schools Extension to 29 schools Optional for all primary schools	none	cohorts 2011,2016
Lower-Saxony	1994-2002 2003	Model Projects in 10 schools Optional for all primary schools	none	cohorts 2006, 2011,2016
Mecklenburg-Vorpommern	2005-2007 2019	Model Projects in 16 schools Optional in all primary schools	none	none
North Rhine-Westphalia	1999-2004	Model Projects in 6 schools	none	none
Rhineland-Palatinate	1995-1998	Model Projects in 2 schools Gradual Extension to 20 schools	none	none
Saarland		no flexible entrance reforms	none	none
Saxony	2001-2004	Model Projects in 25 schools	none	none
Saxony-Anhalt	1997-2000 2000	Model Projects in 4 schools State-wide implementation	cohorts 2006, 2011,2016	cohorts 2006, 2011,2016
Schleswig-Holstein	1994-1997	Model Projects in 5 schools Gradual Extension to 12 schools	none	none
Thuringia	1997 1999-2003 2003-2008 Since 2008	Optional in all primary schools Model Projects in 14 schools Transfer Projects in 25 schools Gradually region-wide implementation	cohort 2016	cohort 2016

Notes: Own collection of information in legal documents and websites of the states' education ministries. The fourth and fifth columns indicate whether fourthgraders of the respective cohorts in the given state are part of a mandatory resp. optional flexible school entrance system. The category "optional" includes both mandatory and optional flexible school entrance rules.

Saxony-Anhalt, in 2000, was the first state to introduce a mandatory flexible entry stage. Berlin followed in 2005 with a state-wide mandatory implementation, but reintroduced the choice between the traditional system and the multigrade approach in 2010. Thuringia made the flexible school entrance mandatory from 2008 onwards and Brandenburg followed in 2010. All of the other states did not introduce a mandatory system of early multigrade teaching. However, some of them made it optional for schools to establish a flexible school entrance stage: Bremen, Hesse, Lower-Saxony, and Berlin. The rest of the 16 states decided - after experimenting in some pilot schools - against a broader implementation of the flexible school entrance.

2.3 Data and Empirical Strategy

2.3.1 Data

Our analysis combines two data sources, which enables us to produce a longitudinal dataset that is representative for fourthgraders' performance and motivation in German primary schools. The two data sources are the *Progress in International Reading Literacy Study* (PIRLS) in 2001 and 2006, and the National Assessment Study in 2011 and 2016 by the *German Institut zur Qualitätsentwicklung im Bildungswesen* (IQB) (Institute of Quality Development in the Education System). Both data sources have in common that they provide state identifiers. Those are necessary for the linkage with our reform data.[6] The first source, the extended PIRLS assessment in Germany, not only includes reading test scores, but also students' grades in German, the recommendation for the next school track, as well as students' school enjoyment. We make also use of the information available on student and family background in order to control for factors that may impact students' education outcomes. In a robustness check, we also use available teacher and school characteristics as control variables. The second source, the German National Assessment Study, also assesses reading test scores of fourth-graders, comparable to the PIRLS and takes place at the end of primary school. It includes also information on the other outcomes of interest as well as the control variables in the same way as the PIRLS data. Our final sample thus comprises students in their fourth grade in 2001, 2006, 2011, or 2016; and who entered their first school year in 1998, 2003, 2008, or 2013 respectively. The combined data yields a sample of approximately 68,000 students.

[6] Note that the PIRLS data for the years 2011 and 2016 do not include state identifiers anymore. Therefore, we have to rely on the IQB data. Since the IQB studies are very similar to the design of PIRLS, however, the combination of both data sources is possible.

2.3.2 Empirical Model

The combination of the different reforms in the various states with cross-sectional data of outcomes for four cohorts of fourth-graders (2001, 2006, 2011, and 2016) implies the following research design: a cohort of fourth-graders was part of a flexible school entrance system if the reform had been in place when the cohort entered first grade. For example, since Saxony-Anhalt introduced a mandatory school entrance stage in 2000, the cohort 2001 was not treated by the reform yet, but the cohorts 2006, 2011, and 2016 were treated. This is shown in the fourth column of Table 2.1 for every state. The fifth column shows the affected cohorts when we are not only considering mandatory flexible school entrance systems, but in addition all states that made it optional for schools. We use the *mandatory* definition for our main analysis of the effects of the flexible school entrance system. However, we use the *optional* definition in robustness checks.

We use the staggered implementation of the flexible entrance stage across German states to estimate the effect of multigrade classrooms in a difference-in-differences framework. This approach exploits the variation in the exposure to a multigrade class (i) across reforming and non-reforming states and (ii) between affected and unaffected cohorts within the same federal state. Our main specification is thus given as follows:

$$Y_{i,s,t} = \beta_0 + \beta_1 multigrade_{s,t} + X_i\beta_2 + \mu_s + \mu_t + \epsilon_{i,s,t}. \tag{2.1}$$

where $Y_{i,s,t}$ is the outcome variable for student i in cohort t attending school in state s. The dummy variable *multigrade* equals 1 for the treated states in the treatment period. In our baseline analysis we only define those states as treated if they introduced a *mandatory* flexible school entrance stage. This has the advantage that all students of a cohort who got enrolled in primary school during a treatment period were certainly experiencing a multigrade setting during their first school years. The disadvantage is that the observed students in the control states could have been also treated if their states had an optional rule regarding the flexible school entry stage (see Table 2.1). This implies a mis-classification of treatment and control and might lead to a bias in the estimates towards zero. Therefore, as a robustness check we define all states as being treated which introduced mandatory or *optional* multigrade classes. Since, however this latter definition does imply that very probably not all students in the treatment group are actually treated, it rather has to be interpreted as an intention-to-treat effect.

The vector X_i includes a set of control variables to account for students' demographic characteristics. Our baseline analysis controls for gender and age, kindergarten attendance, migrant background, and parental education. In a robustness check, we control for books at home instead of parental education. In further robustness checks, we additionally control for teacher and school characteristics.

State fixed effects μ_s control for time-invariant conditions in each state, including state capacity, local culture, or geography. Cohort fixed effects μ_t capture national trends in student cohorts' demographic composition, as well as general trends in the education sector or the labor market. $\epsilon_{i,s,t}$ is the error term. We cluster the standard errors at the state level as the treatment varies as the state level. Considering recent developments in the econometric literature we calculate p-values of two different types of clustering methods for each reform coefficient displayed in our tables. First, we use the standard clustering method which is conservative in our kind of setting and accounts for potential correlation of error terms across years within states (Athey and Imbens, 2018). Second, to account for the limited number of clusters (because there are only 16 German states) we calculate wild cluster bootstrap p-values (Roodman et al., 2019).

Under the assumptions of the difference-in-differences framework, the coefficient β_1 represents the causal effect of the reform. Most importantly, the common trend assumption implies that - in absence of the treatment - reforming and non-reforming states would both lie on the same trend with respect to outcome variables. It is typically argued that this assumption is likely to be fulfilled if the pre-trends prior to the reform are the same in reforming and non-reforming states. Since our data only covers four points in time it is not possible to investigate pre-trends of the outcomes. A specific feature of our main analysis is that only states in East Germany introduced mandatory flexible school entry stages. We therefore, in a further specification, restrict our sample to only East German states. This makes it even more likely that control and treatment states have common trends. Interestingly (and reassuringly), the results do not differ much from the main specification.

A second crucial assumption of our identification strategy is that the treatment effect does not represent any development simultaneously occurring to the multigrade reforms. To avoid this problem, we investigate whether other education reforms affecting primary school students were simultaneously introduced. Indeed, a reform abolishing numerical grades in the first years of primary school has been introduced in some of the states during a similar time frame, yet with a different timing pattern across states. We test the robustness of our results by controlling for the early grading reform, and show that our results are not affected.

As described in Section 1.3, a major reason for the introduction of the reform was the heterogeneous school readiness of children at the beginning of primary school. If children with a lower school readiness stayed longer in kindergarten before the reform, but reduced time in kindergarten after the reform due to the integrative approach of the flexible school entrance level, this would threaten our identification strategy. It is well studied in the literature that years in kindergarten have a positive effect on child development and school performance. If the reform reduced years in kindergarten this could lead to a negative result, but the teaching in multigrade classes would not be the cause for it. Therefore, we use years in kindergarten as placebo outcome. We find that the reform did not affect time in kindergarten.

Finally, since years in kindergarten could lead to being better prepared for following a multigrade class, we interact the reform with kindergarten years, and indeed find heterogeneous effects.

To explore whether girls and children from a low socio-economic background are affected in different ways by being taught in a multigrade classroom, we perform separate analyses for these subgroups and study heterogeneous effects by gender and by parental education.

2.3.3 Descriptives

Table 2.2 shows the descriptive statistics of our dataset.

Reform Variables. As described in Section 1.3 not all reforming states made the flexible school entrance stage, which introduced multigrade classes, a mandatory policy for their schools. Linking the reform data from Table 2.1 to the students observed in our data, it shows that 12% of students experienced a system of a mandatory flexible school entrance stage, see Panel A of Table 2.2. In our main analysis we use this definition of being treated by the reform. The second row of Table 2.2 shows that our alternative definition of the treatment, i.e. being treated if the state has introduced an optional or mandatory flexible school entrance stage, leads to 31% of students belonging to the treatment group.

Outcome Variables. *Reading test scores.* Students' reading test scores are measured by the standardized reading tests provided by the PIRLS resp. IQB study. The test scores from all datasets are originally constructed to have a mean of 500 and standard deviation of 100, thereby facilitating nation-wide comparison. They are z-standardized for the purpose of this study (see first row of Panel B of Table 2.2).

Grades We use the information on the last grade student received for their performance in German. They are graded according to the German grading scale, which varies from 1 (*excellent, sehr gut*) to 6 (*insufficient, ungenügend*). We inverse the scale for the purpose of readability. The second row of Panel B of Table 2.2 shows that students in our dataset receive on average a grade between "good" and "satisfactory".

Recommendation for Gymnasium. As described in Section 1.3, in Germany students are tracked into three differents tracks after primary school. In fourth grade, they receive a

Table 2.2: Descriptive Statistics

	Mean	St.Dev	Min	Max	Observations
Panel A: Reform					
Mandatory Multi-grade	0.12	0.32	0.00	1.00	72873
Optional Multi-grade	0.31	0.46	0.00	1.00	72873
Panel B: Outcomes					
Standardized Reading Testscore	0.00	1.00	-4.46	3.67	68453
German Grade (1=lowest, 6=highest)	4.47	0.90	1.00	6.00	63953
Recommendation for Gymnasium (Dummy)	0.34	0.47	0.00	1.00	69345
Enjoy School (Dummy)	0.68	0.46	0.00	1.00	55009
Panel C: Student Controls					
Student is a girl	0.49	0.50	0.00	1.00	72380
Age of student (in years)	10.47	0.50	6.42	12.92	72346
Low parental education (Dummy)	0.20	0.40	0.00	1.00	72873
Books at Home (1=(<10) to 5=(>200))	3.33	1.20	1.00	5.00	63233
First generation migrant	0.06	0.23	0.00	1.00	72873
Years spent in kindergarten	3.30	0.97	0.00	5.50	72873
Panel D: Teacher Controls					
Age of teacher (in years)	46.91	10.31	24.00	72.00	63578
Experience of teacher (in years)	20.83	12.31	0.00	57.00	63754
Teacher specialized in German (Dummy)	0.81	0.39	0.00	1.00	64116
Teacher works full-time (Dummy)	0.72	0.45	0.00	1.00	64456
Panel E: School Controls					
No. of students enrolled in school	276.48	151.35	12.00	2008.00	67490
Public School (yes/no)	0.96	0.20	0.00	1.00	69412
Experience as headmaster in this school (in years)	9.14	7.12	0.00	42.00	65712
Age of headmaster in years (in years)	52.34	7.42	22.00	71.00	66223
Headmaster is male (Dummy)	0.32	0.47	0.00	1.00	67978

Notes: The table shows the descriptive statistics of a quasi-panel of fourthgraders using data from the PIRLS assessment and the German National assessment (IQB) for the years 2001, 2006, 2011, and 2016.

recommendation by their teacher on which track would be most suitable given the student's ability. We use this information to create a dummy variable indicating whether a student is recommended to enroll in the highest track (Gymnasium). The data show that a third of students in our sample receive this recommendation.

Enjoy going to school. Students were asked to what extent they agree that going to school is enjoyable for them. The answers include the four categories "strongly agree", "somewhat agree", "neither agree nor disagree", and "strongly disagree". We create a dummy variable for enjoying going to school, which takes value 1 if the student strongly or somewhat agrees, and 0 otherwise.

Control Variables. *Student Controls.* Panel C of Table 2.2 shows the individual controls we use in our main analysis. In our dataset, 49% of students are female. The average age is 10.47 years, which is the usual age of fourthgraders. 20% of students have low educated parents, i.e. their parents have at most a lower secondary degree. As an alternative measure for socio-economic background we use the number of books at home as proxy for parental educational background in a robustness check. On average, children's families have a bit more than 100 books at home (as category 3 contains 26-100 books, and category 4 contains 101-200 books). 6% of the students are first generation immigrants. On average, the students have spent 3.3 years in kindergarten prior to primary school.

Teacher Controls. Panel D of Table 2.2 shows teacher characteristics which we use as controls in a robustness check. On average, teachers are 46 years old and have 20 years of experience in the teaching profession. 81% of teachers are specialized in teaching German, and 72% of them work full-time.

School Controls. Finally, Panel E of Table 2.2 displays the descriptive statistics of school characteristics. As in case of the teacher controls, we use these as control variables in a robustness check. On average, the primary schools under study have 276 enrolled students. 96% of the schools are public schools (note that private schools are uncommon in Germany). The headmasters of the respective schools are on average 52 years old, have 9 years of experience as headmaster, and 32% of them are male.

2.4 Results

2.4.1 Main Results

Table 2.3 shows the results of estimating Equation 1 with reading test scores of fourthgraders as outcome variable. In each column, we add one of the individual control variables. The specification in column (1) only uses state and cohort fixed effects. The multigrade coefficient is negative, but small and not significant. Column (2) adds gender and age as controls, which leads to a larger multigrade coefficient, which is still not significant, however. Interestingly, when adding years spent in kindergarten as control variable in column (3), the multigrade coefficient is significant at the 1%-level and equals -7.6% of a standard deviation. Adding a control for being a first generation immigrant in column (4) and having low-educated parents in column (5) leaves the multigrade coefficient significant and economically meaningful. According to the estimation results in column (5), being in a cohort which experienced reform-induced multigrade teaching in the first years of primary school leads to a decline in reading test scores of 6.1% of a standard deviation. The specification of model (5) serves as our main specification in the next steps of the analysis as it has the highest explanatory power (measured by R^2).

Table 2.3: Effect of Multigrade Class on Reading Test Scores of Fourth Graders

	Reading Skills				
	(1)	(2)	(3)	(4)	(5)
Multigrade	-0.019	-0.041	-0.076***	-0.094***	-0.061***
	(0.029)	(0.040)	(0.017)	(0.024)	(0.020)
Female		0.169***	0.160***	0.157***	0.163***
		(0.010)	(0.010)	(0.010)	(0.010)
Age		-0.446***	-0.410***	-0.388***	-0.354***
		(0.019)	(0.016)	(0.017)	(0.016)
KiGa Attended (years)			0.130***	0.117***	0.098***
			(0.014)	(0.013)	(0.011)
Migration Background				-0.337***	-0.342***
				(0.026)	(0.024)
Low SES					-0.430***
					(0.029)
State FE	Yes	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes	Yes
R-squared	0.077	0.135	0.167	0.180	0.203
WCB P-Value	0.625	0.638	0.072	0.145	0.153
Observations	68,453	68,453	68,453	68,453	68,453

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). Reading testscores are z-standardized. KiGa means "*kindergarten*". Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Table 2.4 shows the main effects of early grading on two other achievement measures – the most recent grade in German and the recommendation for the high track, as well as on the motivational outcome – measured as enjoying school. Column (1) displays the negative effect on reading testscores described above. Column (2) shows that the multigrade reform had also a significant negative effect on the grade in German which equals approximately 1/9 of a standard deviation (0.108 divided by the sample standard deviation 0.9, see Table 2.2). Despite the negative impact on both test scores and grades, neither the high-track recommendation nor enjoyment of school are significantly affected by the multigrade reform, even though the coefficients in columns (3) and (4) are also negative.

Table 2.4: Effect of Multigrade Class on Further Outcomes of Fourth Graders

	Reading Test Score (1)	German Grade (2)	High-Track Recomm (3)	Enjoy School (4)
Multigrade	-0.061***	-0.108***	-0.046	-0.019
	(0.020)	(0.034)	(0.038)	(0.011)
Female	0.163***	0.258***	0.029***	0.151***
	(0.010)	(0.010)	(0.004)	(0.006)
Age	-0.354***	-0.391***	-0.144***	-0.010**
	(0.016)	(0.010)	(0.012)	(0.004)
Migration Background	-0.342***	-0.263***	-0.076***	0.023***
	(0.024)	(0.020)	(0.014)	(0.006)
KiGA Attended (years)	0.098***	0.081***	0.038***	0.000
	(0.011)	(0.006)	(0.005)	(0.003)
Low SES	-0.430***	-0.397***	-0.204***	-0.024**
	(0.029)	(0.016)	(0.016)	(0.008)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.203	0.195	0.138	0.044
WCB P-Value	0.167	0.059	0.699	0.376
Observations	68,453	63,953	69,345	55,009

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). Reading testscores are z-standardized. Grades in German are on a scale from 1 (”insufficient”) to 6 (”very good”). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school.Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

2.4.2 Effect Heterogeneity

Boys & Girls

Earlier findings by Leuven and Rønning (2016) and Gerhardts et al. (2021a) indicate heterogeneity of effects of multigrade teaching by gender and parental education. The findings of Gerhardts et al. (2021a) suggest that the negative effect of multi grade classes is stronger for girls than for boys, and document a more pronounced negative effect on children of blue-collar workers.

Table 2.5 shows that girls are significantly negatively affected by multigrade classes in terms of their reading test scores (-0.08), their grades in German (-0.123) and their enjoyment of school (-0.025). Boys, on the contrary, do not seem to be as harmed by being taught in a multigrade classroom. The effect on their reading test scores are smaller

(-0.04) and not significant, and whether they enjoy going to school is also not affected. Boys' grades in German, however, are significantly affected, but less than in the case of girls (-0.092).

Consequently, these subgroup results are in line with the evidence of our study on the reforms in the Saarland several decades before.

Table 2.5: Effect of Multigrade Class on Girls

	Reading Test Score	German Grade	High-Track Recomm	Enjoy School
	(1)	(2)	(3)	(4)
Multigrade	-0.080***	-0.123***	-0.057	-0.025*
	(0.026)	(0.028)	(0.038)	(0.013)
Age	-0.358***	-0.384***	-0.141***	-0.018**
	(0.018)	(0.012)	(0.013)	(0.007)
Migration Background	-0.319***	-0.234***	-0.062***	0.008
	(0.028)	(0.027)	(0.017)	(0.013)
KiGa Attended (years)	0.096***	0.079***	0.037***	-0.001
	(0.011)	(0.007)	(0.005)	(0.003)
Low SES	-0.442***	-0.417***	-0.218***	-0.024**
	(0.026)	(0.021)	(0.017)	(0.010)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.188	0.174	0.141	0.024
WCB P-Value (Reform)	0.089	0.055	0.537	0.333
N	33,763	31,541	34,144	27,363

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of female 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). Reading testscores are z-standardized. Grades in German are on a scale from 1 (”insufficient”) to 6 (”very good”). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school. KiGa means “*kindergarten*”. Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Parental Education Background

Table 2.A.1 in the Appendix shows that children with high parental education are significantly negatively affected by multigrade classes in terms of their reading test scores (-0.088) as well as their grades in German (-0.123). Surprisingly, children with low parental education are not significantly affected by the multigrade reform, the reform coefficient is negative but rather small (-0.021), see Table 2.A.2 in the Appendix. Their grades in

Table 2.6: Effect of Multigrade Class on Boys

	Reading Test Score	German Grade	High-Track Recomm	Enjoy Reading
	(1)	(2)	(3)	(4)
Multigrade	-0.040	-0.092**	-0.036	-0.014
	(0.035)	(0.042)	(0.039)	(0.019)
Age	-0.352***	-0.397***	-0.146***	-0.002
	(0.017)	(0.013)	(0.012)	(0.005)
Migration Background	-0.364***	-0.287***	-0.089***	0.035***
	(0.030)	(0.020)	(0.014)	(0.008)
KiGa Attended (years)	0.100***	0.082***	0.039***	0.002
	(0.013)	(0.007)	(0.004)	(0.004)
Low SES	-0.420***	-0.376***	-0.189***	-0.026**
	(0.036)	(0.019)	(0.015)	(0.010)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.204	0.174	0.132	0.016
WCB P-Value (Reform)	0.545	0.005	0.801	0.569
N	34,690	32,412	35,201	27,646

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of male 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). Reading testscores are z-standardized. Grades in German are on a scale from 1 (”insufficient”) to 6 (”very good”). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school. KiGa means “*kindergarten*”. Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

German, however, are significantly affected, but less than in the case of children with high-educated parents (-0.091). There is no significant effect on the high-track recommendation or the enjoyment of school for neither of both groups.

Finding worse results for children with more advantaged family backgrounds is in contrast to our findings on the effects of the reform in the Saarland (Gerhardts et al., 2021a). In some way, the result indicates that educational inequality could decrease due to the multigrade classes. Unfortunately, this seems to come at the cost of deteriorating skills of the more advantaged group of students rather than through more enhanced skills of the disadvantaged group.

2.5 Robustness Checks

2.5.1 Years in Kindergarten as Placebo Outcome

In this section, we test the robustness of our main results presented in Section 2.4 and discuss some of the assumptions explained in Section 1.4 in more depth.

A major reason for the introduction of the reform was the heterogeneous school readiness of children at the beginning of primary school. If children with a lower school readiness stayed longer in kindergarten before the reform, but reduced time in kindergarten after the reform due to the integrative approach of the flexible school entrance level, this would threaten our identification strategy. Therefore, we use years in kindergarten as placebo outcome. Table 2.A.3 in the Appendix shows that there is no effect of the reform on time spent in kindergarten. Looking at the control variables, the familiar pattern of socio-economic selection into kindergarten is visible. Both migrant background as well as low parental education are negatively associated with the intensive margin of kindergarten attendance.

2.5.2 Interaction of Reform with Years in Kindergarten

As a longer preparation for school children receive in kindergarten could enable them to cope with the situation in multigrade classes better, we investigate the interaction of the reform with years spent in kindergarten. Table 2.A.4 in the appendix shows that the interaction is significantly positive. This implies that spending more years in kindergarten before primary school mitigates the negative effect of being taught in a multigrade classroom. Interestingly, adding the interaction shows that children who spend less time in kindergarten not only experience significant negative effects in terms of their test scores and grade, but also in terms of their high-track recommendation and school enjoyment (columns (3) and (4)).

2.5.3 Controlling for Another Reform in Primary Schools

A second crucial assumption of our identification strategy is that the treatment effect does not represent any development simultaneously occurring to the multigrade reforms. To avoid this problem, we investigate whether other education reforms affecting primary school students were simultaneously introduced. Indeed, a reform abolishing numerical grades in the first years of primary school has been introduced in four of the states during a similar time frame, yet with a different timing pattern across states (Hesse in 1999, Saarland in 2000, Brandenburg in 2001, Berlin in 2006). We test the robustness of our results by controlling for the early grading reform in Table 2.A.5 in the Appendix. The table shows that our results are not affected.

2.5.4 Sample Restricted to East German States

A specific feature of our main analysis is that only states in East Germany introduced mandatory flexible school entry stages. We therefore, in a further specification, restrict our sample to only East German states. This makes it even more likely that control and treatment states have common trends. Table 2.7 shows that the results are robust and do not differ much from the main specification.

Table 2.7: Effect of Multigrade Class on Students' Outcomes - East Germany

	Reading Test Score	German Grade	High-Track Recomm	Enjoy School
	(1)	(2)	(3)	(4)
Multigrade	-0.057**	-0.061***	-0.029	-0.018
	(0.020)	(0.013)	(0.017)	(0.010)
Female	0.196***	0.292***	0.027**	0.146***
	(0.013)	(0.015)	(0.009)	(0.008)
Age	-0.382***	-0.386***	-0.122***	-0.015
	(0.029)	(0.019)	(0.023)	(0.008)
Migration Background	-0.288***	-0.216***	-0.026	0.040***
	(0.041)	(0.038)	(0.023)	(0.009)
KiGa Attended (years)	0.082***	0.087***	0.027**	0.006
	(0.018)	(0.009)	(0.007)	(0.004)
Low SES	-0.411***	-0.349***	-0.171***	-0.039**
	(0.083)	(0.033)	(0.024)	(0.015)
Student Controls	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.186	0.180	0.137	0.044
WCB P-Value	0.320	0.011	0.443	0.366
Observations	24,366	24,681	25,177	21,094

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of only East German 4th-grade students using data from the PIRLS assessment and a German National assessment (IQB). Reading scores are z-standardized. KiGa means "*kindergarten*". Standard errors are clustered at the state level. Grades are on a scale from 1 ("insufficient") to 6 ("very good"). High track is a dummy being one if teacher recommends high track. Enjoy school is a dummy equaling one if student fully agrees to enjoy going to school. Different number of observations due to differing availability of outcome variable. * denotes statistical significance at the 10% level, ** at the 5% level and *** at the 1% level.

2.5.5 Parental Background Control

Family background is often measured by parental education, as we do in our main specification. However, many studies have shown that also the variable "books at home" is a very reliable proxy for the socio-economic status of a family. Therefore, in a robustness check, we control for this variable instead of parental education. The results of both specifications are very similar, as Table 2.A.6 in the Appendix shows.

2.5.6 Teacher and School Characteristics as Further Controls

To alleviate the assumption of common trends of treated and control states a little bit we add teacher and school characteristics as control variables. If the composition of teachers or organizational patterns of the schools changed along with the multigrade reforms, adding these controls would make a difference for our estimates. Table 2.A.7 in the Appendix shows no important differences in comparison to our main specification, however. In addition, the explanatory power (R2) does not increase much by adding these further controls.

2.5.7 Definition of Treatment Status

Finally, we check the robustness of our results with respect to the definition of the treatment status of the cohorts in our sample. In our baseline analysis we only define those states as treated which introduced a *mandatory* flexible school entrance stage. This has the advantage that all students of a cohort who got enrolled in primary school during a treatment period were certainly experiencing a multigrade setting during their first school years. The disadvantage is that the observed students in the control states could have been also treated if their states had an optional rule regarding the flexible school entry stage (see Table 2.1) and they happen to be in a school that makes use of this option.[7] This is likely to lead to a downward bias in our estimates. Therefore, as a robustness check we define all states as being treated which introduced mandatory or *optional* multigrade classes. Table 2.8 shows that the effects on reading test scores and grades stay significant using the new definition, the coefficients are (in absolute terms) larger which is in line with the argument stated above – moving the states with optional flexible school entrance systems to the treatment group removes all potentially treated observations out of the control group.

[7] Note that the assignment to primary schools is based on school catchment areas in Germany, so that sorting to schools dependent on whether they introduced a flexible school entrance stage is not possible.

Table 2.8: Effect of Optional Multigrade Class on Students' Outcomes

	Reading Test Score (1)	German Grade (2)	High-Track Recomm (3)	Enjoy School (4)
Optional Multigrade	-0.086*	-0.100*	-0.044	-0.023
	(0.047)	(0.050)	(0.040)	(0.019)
Female	0.163***	0.258***	0.029***	0.151***
	(0.010)	(0.010)	(0.004)	(0.006)
Age	-0.354***	-0.389***	-0.143***	-0.010**
	(0.016)	(0.010)	(0.012)	(0.004)
Migration Background	-0.342***	-0.264***	-0.076***	0.022***
	(0.024)	(0.021)	(0.014)	(0.006)
KiGA Attended (years)	0.099***	0.081***	0.039***	0.000
	(0.012)	(0.007)	(0.005)	(0.003)
Low SES	-0.430***	-0.398***	-0.204***	-0.024***
	(0.030)	(0.015)	(0.015)	(0.008)
Student Controls	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.203	0.195	0.138	0.044
WCB P-Value	0.125	0.101	0.285	0.295
Observations	68,453	63,953	69,345	55,009

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and a German National assessment (IQB). Treatment is going to school in a state which introduced mandatory *or* optional multigrade classes. Reading scores are z-standardized. Standard errors are clustered at the state level. Grades are on a scale from 1 ("insufficient") to 6 ("very good"). High track is a dummy being one if teacher recommends high track. Enjoy school is a dummy equaling one if student fully agrees to enjoy going to school. Different number of observations due to differing availability of outcome variable. * denotes statistical significance at the 10% level, ** at the 5% level and *** at the 1% level.

2.6 Conclusion

This paper provides novel evidence about the impact of exposure to multigrade teaching in primary school on educational outcomes. The results of a difference-in-differences approach that exploits the staggered implementation of flexible school entrance levels across German states between 1997 and 2010 reveal a significant negative effect of multigrade teaching on educational outcomes such as reading test scores and grades, but no effect on teacher recommendations or subjective perceptions of pupils. This partly rationalizes the mixed evidence in the literature by documenting that multigrade teaching does not exhibit negative effects throughout. Instead, the effects emerge for skills that can be measured in comparable metrics. The effects are more pronounced for girls, complementing earlier evidence from other studies in different contexts. The evidence also shows that spending more years in kindergarten before primary school mitigates the negative effects of exposure to multigrade teaching.

In light of these findings, more work is needed to reveal the mechanisms underlying these effects.

2.A Appendix

Table 2.A.1: Effect of Multigrade Class on Children with High Parental Education

	Reading Test Score	German Grade	High-Track Recomm	Enjoy Reading
	(1)	(2)	(3)	(4)
Multigrade	-0.088***	-0.123***	-0.048	-0.015
	(0.028)	(0.040)	(0.036)	(0.011)
Female	0.175***	0.269***	0.033***	0.150***
	(0.011)	(0.009)	(0.005)	(0.007)
Age	-0.370***	-0.398***	-0.150***	-0.008
	(0.017)	(0.012)	(0.013)	(0.005)
Migration Background	-0.367***	-0.294***	-0.092***	0.024***
	(0.026)	(0.022)	(0.016)	(0.007)
KiGA Attended (years)	0.105***	0.083***	0.043***	0.002
	(0.012)	(0.006)	(0.005)	(0.003)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.188	0.195	0.142	0.041
WCB P-Value (Reform)	0.163	0.056	0.677	0.410
N	53,927	50,407	55,197	43,572

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students with high-educated parents using data from the PIRLS assessment and the German National assessment (IQB). Reading testscores are z-standardized. Grades in German are on a scale from 1 (”insufficient”) to 6 (”very good”). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school.Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Table 2.A.2: Effect of Multigrade Class on on Children with Low Parental Education

	Reading Test Score (1)	German Grade (2)	High-Track Recomm (3)	Enjoy Reading (4)
Multigrade	-0.021	-0.091***	-0.011	-0.011
	(0.040)	(0.023)	(0.028)	(0.019)
Female	0.121***	0.216***	0.011	0.157***
	(0.015)	(0.020)	(0.007)	(0.010)
Age	-0.293***	-0.359***	-0.120***	-0.017
	(0.014)	(0.015)	(0.012)	(0.011)
Migration Background	-0.247***	-0.147***	-0.018	0.011
	(0.029)	(0.032)	(0.014)	(0.017)
KiGa Attended (years)	0.071***	0.072***	0.028***	-0.004
	(0.010)	(0.006)	(0.006)	(0.006)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.283	0.146	0.078	0.055
WCB P-Value (Reform)	0.596	0.042	0.720	0.571
N	14,526	13,546	14,148	11,437

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students with low-educated parents using data from the PIRLS assessment and the German National assessment (IQB). Reading test scores are z-standardized. Grades in German are on a scale from 1 (”insufficient”) to 6 (”very good”). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school. KiGa means “*kindergarten*”. Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Table 2.A.3: Effect of Multigrade Class on Placebo Outcome: Kindergarten Attendance

	Kindergarten Attendance				
	(1)	(2)	(3)	(4)	(5)
Multigrade	0.152	0.149	0.161	0.160	0.173
	(0.237)	(0.238)	(0.242)	(0.243)	(0.239)
Female		-0.030***	-0.021***	-0.023***	-0.019**
		(0.007)	(0.007)	(0.007)	(0.007)
Age		-0.040***	-0.065***	-0.039**	-0.021
		(0.013)	(0.014)	(0.014)	(0.013)
Migration Background				-0.484***	-0.480***
				(0.035)	(0.035)
Low SES					-0.213***
					(0.028)
State FE	Yes	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes	Yes
R-squared	0.274	0.274	0.288	0.300	0.307
WCB P-Value	0.650	0.614	0.610	0.603	0.558
Observations	72,873	72,873	72,873	72,873	72,873

*Notes:*The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). The placebo outcome used here is kindergarten attendance (years). Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Table 2.A.4: Effect of Multigrade Class on Students' Outcomes - Interaction term: Reform and time spent in kindergarten

	Reading Test Score (1)	German Grade (2)	High-Track Recomm (3)	Enjoy School (4)
Multigrade	-0.079**	-0.132***	-0.072*	-0.047**
	(0.035)	(0.014)	(0.039)	(0.016)
Interaction Kindergarten	0.037	0.042*	0.036*	0.034**
	(0.021)	(0.023)	(0.018)	(0.014)
Kindergarten	0.166***	0.124***	0.042***	-0.012**
	(0.018)	(0.010)	(0.009)	(0.005)
Female	0.162***	0.257***	0.028***	0.151***
	(0.010)	(0.010)	(0.004)	(0.006)
Age	-0.358***	-0.394***	-0.145***	-0.010**
	(0.016)	(0.011)	(0.012)	(0.004)
Migration Background	-0.370***	-0.287***	-0.089***	0.022***
	(0.025)	(0.019)	(0.014)	(0.006)
Low SES	-0.440***	-0.405***	-0.209***	-0.025***
	(0.032)	(0.017)	(0.016)	(0.008)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.200	0.192	0.135	0.044
WCB P-Value	0.193	0.140	0.083	0.126
Observations	68,453	63,953	69,345	55,009

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). Reading testscores are z-standardized. Grades in German are on a scale from 1 (”insufficient”) to 6 (”very good”). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school. Kindergarten is a dummy measuring one if a child spent more than 3 years in child care before school. Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Table 2.A.5: Effect of Multigrade Class on Students' Outcomes - Controlling for Early Grading Reform

	Reading Test Score	German Grade	High-Track Recomm	Enjoy School
	(1)	(2)	(3)	(4)
Multigrade	-0.060***	-0.108***	-0.046	-0.020*
	(0.016)	(0.036)	(0.037)	(0.011)
Early Grading	0.054	-0.006	0.016	-0.020
	(0.043)	(0.083)	(0.038)	(0.034)
Female	0.163***	0.258***	0.029***	0.151***
	(0.010)	(0.010)	(0.004)	(0.006)
Age	-0.355***	-0.391***	-0.144***	-0.009*
	(0.016)	(0.011)	(0.012)	(0.004)
Migration Background	-0.342***	-0.263***	-0.076***	0.023***
	(0.024)	(0.020)	(0.014)	(0.006)
KiGa Attended (years)	0.099***	0.081***	0.038***	0.000
	(0.012)	(0.006)	(0.005)	(0.003)
Low SES	-0.430***	-0.397***	-0.204***	-0.024**
	(0.029)	(0.016)	(0.016)	(0.008)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.203	0.195	0.138	0.044
WCB P-Value	0.136	0.083	0.729	0.491
Observations	68,453	63,953	69,345	55,009

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). In addition to the main specification, we control for a reform which introduced early numerical grading in some of the German states between 1999 and 2006. Reading testscores are z-standardized. Grades in German are on a scale from 1 (”insufficient”) to 6 (”very good”). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school. KiGa means “*kindergarten*”. Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Table 2.A.6: Effect of Multigrade Class on Students' Outcomes - "Books at home" as Parental Background Control

	Reading Test Score	German Grade	High-Track Recomm	Enjoy School
	(1)	(2)	(3)	(4)
Multigrade	-0.082***	-0.120***	-0.053	-0.021*
	(0.018)	(0.032)	(0.037)	(0.011)
Female	0.155***	0.244***	0.025***	0.152***
	(0.009)	(0.009)	(0.004)	(0.006)
Age	-0.281***	-0.341***	-0.128***	-0.006
	(0.014)	(0.012)	(0.012)	(0.005)
Migration Background	-0.239***	-0.186***	-0.035**	0.028***
	(0.025)	(0.022)	(0.013)	(0.007)
KiGa Attended (years)	0.071***	0.066***	0.031***	-0.002
	(0.009)	(0.005)	(0.004)	(0.003)
Books at home	0.249***	0.192***	0.092***	0.015***
	(0.011)	(0.006)	(0.006)	(0.002)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.245	0.213	0.160	0.046
WCB P-Value	0.103	0.095	0.545	0.371
Observations	61,071	56,828	60,935	52,452

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). In contrast to the main specification, we use "books at home" instead of parental education as proxy for socio-economic status of the family. Reading testscores are z-standardized. Grades in German are on a scale from 1 ("insufficient") to 6 ("very good"). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school. KiGa means "*kindergarten*". Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Table 2.A.7: Effect of Multigrade Class on Students' Outcomes - Teacher and Schools Characteristics as Controls

	Reading Test Score	German Grade	High-Track Recomm	Enjoy School
	(1)	(2)	(3)	(4)
Multigrade	-0.101***	-0.107***	-0.045	-0.017
	(0.019)	(0.029)	(0.039)	(0.012)
Female	0.162***	0.258***	0.029***	0.151***
	(0.010)	(0.010)	(0.004)	(0.006)
Age	-0.339***	-0.383***	-0.141***	-0.011**
	(0.014)	(0.011)	(0.012)	(0.004)
Migration Background	-0.294***	-0.243***	-0.070***	0.017**
	(0.027)	(0.024)	(0.014)	(0.006)
KiGa Attended (years)	0.085***	0.077***	0.037***	0.002
	(0.009)	(0.005)	(0.004)	(0.003)
Low SES	-0.393***	-0.380***	-0.197***	-0.028***
	(0.020)	(0.013)	(0.016)	(0.008)
Teacher is female	0.025	-0.010	-0.003	0.022**
	(0.026)	(0.023)	(0.012)	(0.010)
Age of teacher	0.002	-0.001*	0.000	-0.001*
	(0.002)	(0.001)	(0.001)	(0.001)
Experience of teacher	-0.001	-0.001	-0.001*	0.001**
	(0.002)	(0.001)	(0.001)	(0.001)
Teacher specialized	0.033*	0.041**	-0.003	0.016**
	(0.017)	(0.017)	(0.008)	(0.007)
Teacher works full-time	-0.019	-0.033**	-0.020**	0.006
	(0.013)	(0.015)	(0.009)	(0.008)
No. of students enrolled	0.000	-0.000	0.000**	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
School is a public School (Dummy)	-0.171***	-0.161***	-0.077***	-0.014
	(0.034)	(0.027)	(0.016)	(0.009)
Experience as headmaster	-0.001	-0.001	-0.000	0.000
	(0.001)	(0.001)	(0.001)	(0.000)
Age of headmaster	0.001	0.001	0.000	-0.000
	(0.001)	(0.001)	(0.001)	(0.000)
Headmaster is male	-0.041**	-0.022	-0.020**	0.010
	(0.019)	(0.015)	(0.009)	(0.007)
State FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.226	0.202	0.143	0.046
WCB P-Value	0.143	0.060	0.773	0.537
Observations	68,453	63,953	69,345	55,009

Notes: The table reports coefficients from a linear regression as given by equation (1) for a quasi-panel of 4th-grade students using data from the PIRLS assessment and the German National assessment (IQB). In contrast to the main specification, we add controls for teacher and school characteristics. Reading testscores are z-standardized. Grades in German are on a scale from 1 (”insufficient”) to 6 (”very good”). Recommendation for Gymnasium is a dummy equal to one if the teacher recommends the high track. Enjoy school is a dummy equaling one if student agrees to enjoy going to school. KiGa means “*kindergarten*”. Standard errors are clustered at the state level. * denotes statistical significance based on the at the 10% level, ** at the 5% level and *** at the 1% level. p-values of the multigrade coefficient using wild cluster bootstrapped t-statistics are displayed at the bottom of the table.

Chapter 3

The Economics of Labor & Patients' Health Outcomes: Evidence from Childbirth in Germany

3.1 Introduction

"Physicians serve the health of the individual and of the population. The medical profession is not a trade. It is by nature a liberal profession." (Model Professional Code for Physicians in Germany, 1997)

"[...]das Handicap ist die moderne Geburtsmedizin, die Geburt und Schwangerschaft zur Risikoaffäre macht." [1]

More than two thousand years ago, physicians declared through the Hippocratic Oath the benefits of the sick to be the sole objective of their profession (Tyson, 2001). In contemporary Germany, the ancient ethical principles are protected by stating medical services to be not for profit. This includes obstetric care, which refers to all treatments related to childbirth.

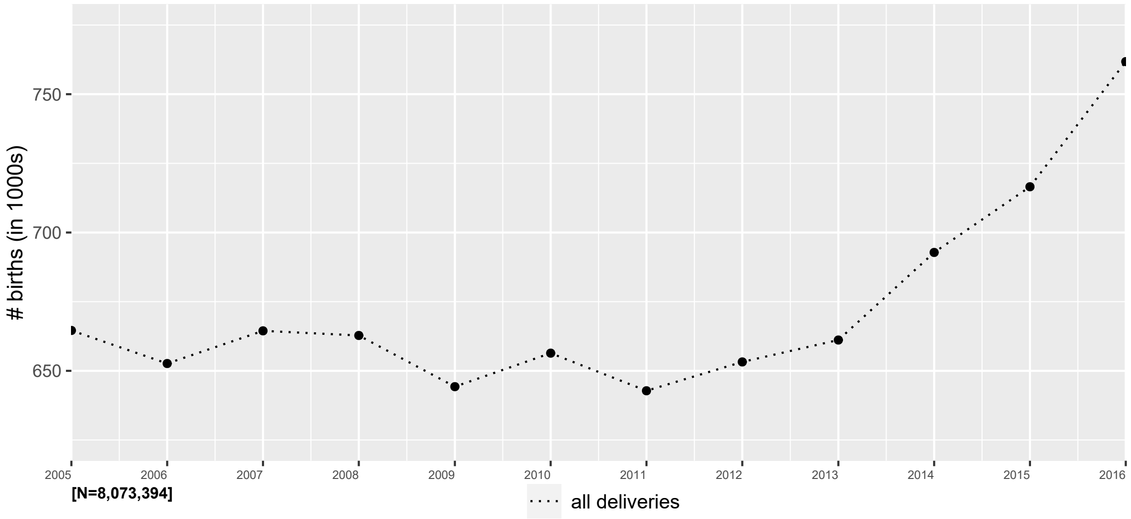
In practice, maternity units operate under pressure by profit-oriented reimbursement schemes paired with acute staff shortages, thus facing adverse incentives for birth interventions (Scharl et al., 2019; Feige, 2008).[2] Because maternity units manage the least predictable hospital events apart from emergency care, they are burdened with substantial non-refundable standby costs for staff-intense patient monitoring (Bruns, 2017). Thereby, hospitals conceding mothers an unassisted vaginal delivery do so at a loss, a dilemma providing incentives for labor induction or surgical birth interventions (Bruns, 2017). As of 2017, 40% of German hospitals provided obstetric care without breaking even (Bruns, 2014), while the share of birth interventions reached twice the size recommended by the WHO (2015).

But what are the consequences of non-medically indicated induced labor for patients' health and a hospital's business operations, in particular, staffing capacities? Laying the ground for compelling causal evidence on the topic, this study applies a novel identification approach to the universe of German hospital births from 2015 and 2016. On the one hand, patients' health impacts are assessed, first and foremost in terms of 1) a severe laceration

[1] *[...] it is modern birth medicine that renders the birth process and the pregnancy risky.* Own translation. Alfred Rockenschaub (2005). Former head of Ignaz Semmelweis Frauenklinik, Vienna. Known for ceasarian section (the surgical delivery through a mother's abdomen, henceforth: c-section) rates about 1% without inflating mortality.

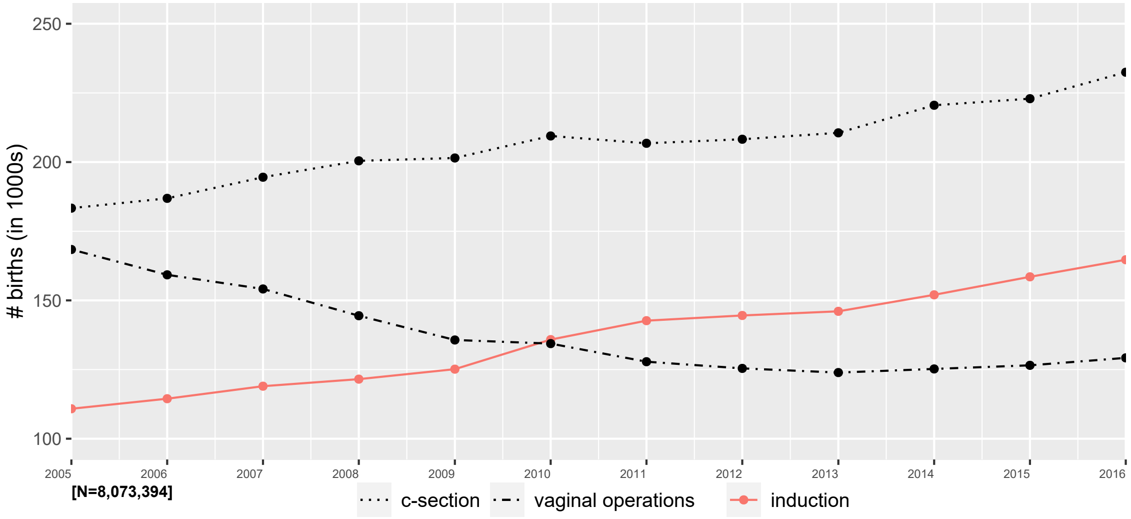
[2] Since 2003, hospitals have been reimbursed based on the Diagnosis-Related-Groups (DRG) system, a flat rate-per-case scheme (Jürges and Köberlein, 2015). There is no (direct) reimbursement for inducing labor (InEK, 2021). Its economic appeal relative to spontaneous labor consists rather in reduced standby costs, i.e., non-refundable costs for staff kept in readiness but not called into action (like a surgeon for an unassisted birth). The rising trends in inductions and other main birth interventions are depicted in Figure 3.2.

Figure 3.1: Total Hospital Admissions for Birth Across Years



Source: IQTIG German hospital birth records. To capture the hospitals' workload in the best possible way, 1) mothers transferred between hospitals are counted repeatedly and 2) still-born neonates are included. E.g., a mother transferred from hospital A to B who delivers a live- and a still-born twin at B is registered once in the records from A and twice in those from B. Own calculations.

Figure 3.2: Absolute Intervention Distribution Across Years



Source: IQTIG German hospital birth records for 2005-2016. Own calculations.

of the mother's birth canal, and 2) the neonatal APGAR score. On the other hand, the effects on a hospital's staffing capacities are primarily captured by 3) labor duration and 4) the postnatal hospital stay.

The rising birth intervention rates have triggered mostly observational evidence for physician-induced demand in the context of childbirth.[3] Very few large-scale causal studies have addressed the medical concerns associating induction with a prolonged and more harmful course of labor as well as adverse health outcomes after birth. Likewise, from an economic perspective, it remains unclear, if and to which extent a health impact of inductions aggravates staff capacity constraints due to additional patient monitoring.

Identifying the impact of non-medically indicated inductions is hard for several reasons. Among the most important ones: Interventions are likely non-randomly assigned and interdependent, e.g., choosing a pre-labor c-section foregoes induction but induction can be followed by c-section. To overcome these challenges, this chapter allows for multiple endogenous treatments. To identify the sole and combined effects of induced labor, c-sections, and vaginal operations six instruments are considered, all of which are new to the health economics literature.

The first three instruments use variation within a given hospital and across obstetricians' preferences to perform a specific intervention. The preference for, e.g., inducing labor corresponds to the mean induction rate across an obstetrician's past deliveries. The fourth instrument exploits if a mother's predicted due date happens to be a working day or not because staff shortages are more acute on non-working days. The fifth instrument exploits if the incidence of a mother experiencing a pre-labor rupture of membranes happens during the night shift because the night shift suffers relatively more from under-staffing than the day shift. The last instrument exploits fluctuations of midwife shortages the moment a mother is admitted to a hospital.

The main findings are twofold. First, induction performed for non-medical reasons strongly impairs patients' health. Second, the adverse health effects imply a staff capacity burden easily overlooked by seminal capacity measures. As to immediate maternal health, induction makes high-degree perineal tearing 6% more likely. Specifically, induction followed by surgical intervention aggravates tearing so much that - given the distribution of single and combined inductions in our main sample - it outweighs the relief in tearing estimated for inductions alone. Besides, severe tearing due to a violent course of labor potentially requires postpartum or later-life surgery (Lydon-Rochelle et al., 2000; Gün et al., 2016; Zahn and Yeomans, 1990). In turn, birth canal surgery is associated with compromising future fertility (Halla et al., 2020; Gizzo et al., 2013; Norberg and Pantano, 2016). As to neonatal health, the detrimental effects (-2.2) found for the APGAR score, the seminal 0-10 range fitness range for newborns, exceed existing quasi-experimental findings, e.g., Lynch et al. (2019) by a factor of ten. Surgical interventions exhibit (weakly) negative health impacts, too.

By contrast, a hospital's staff capacity (measured by labor duration and the postnatal

[3] Table 3.A.1 recaps studies targeting the causal impact of inductions.

hospital stay) is weakly positively affected by induced relative to unassisted birth. Concretely, labor is estimated to shorten by 0.87 hours while a patient's postnatal hospital stay is not significantly impacted at all. In line with intuition, induction-related health compromise should translate into a staff capacity burden. Considering, e.g., just two routine health checks warranted by lower APGAR scores, a tentative back-of-the-envelope calculation suggests extra staffing costs of 11.8 million EUR p.a.[4] Finally, as expected, surgical interventions mechanically shorten labor and prolong mothers' and neonates' postnatal stays ca. 1.5 days. All in all, the labor length relief non-withstanding, the evidence points to negative intervention impacts rebounding (through impaired health) on staff capacity.

This study is the first to incorporate the endogenous and interrelated nature of all three major birth interventions. Thus, it complements the health economics literature in two main ways. First, the impact of induced versus spontaneous labor is cleanly identified. Second, it provides a new benchmark for both, 1) surgical intervention effects identified simultaneously within the same framework, and 2) any birth intervention effect from the literature still relying on single-intervention identification.

Explicitly estimating the impact of induced versus unassisted labor is challenging. By construction, single-treatment identification defaults to comparing induced labor to any other birth mode after some waiting period, so-called expectant management. The few large-scale RCTs report mixed but predominantly positive effects of non-medically indicated induction. However, due to ethical restrictions, they are not blinded and prone to low or selective participation casting doubt on internal and external validity (Carmichael and Snowden, 2019).[5] Besides, autocorrelation of findings arises as trials of multi-side RCTs are referenced repeatedly in systematic reviews (Carmichael and Snowden, 2019).

By contrast, the limited number of large-scale quasi-experimental studies agree on weakly negative effects. Exploiting exogenous shifts in the timing of induction, Buckles and Guldi (2017), Lynch et al. (2019), and Gans and Leigh (2008) find a higher incidence of precipitous labor, birth injuries, etc. Buckles and Guldi (2017), and Jürges (2017) document null effects on c-section likelihood.

Finally, there exists a huge body of mixed observational evidence. If at all, there is some consensus on the detrimental health effects of 1) induced relative to spontaneous labor (Vahratian et al., 2005; Vrouenraets et al., 2005), and 2) induced labor after expectant management relative to induction alone (Harper et al., 2012).

[4] Underlying assumptions and computations are detailed in section 3.5.

[5] The ARRIVE Trial (Grobman et al., 2018), a recent multi-site RCT with a global policy impact, shows a significant decrease in the likelihood of (non-/emergency) c-section (19 vs. 22%), prolonged labor (20 vs. 14 hours), and a slightly shorter maternal postnatal hospital stay. However, only 23% of eligible women participated, with roughly 10% official non-compliers. Anecdotally, Goer (2018) proposes even higher physician-induced non-compliance in the control group.

This study's simultaneous identification and straightforward assessment of the relative impacts of birth interventions fills a gap in the literature. So far, Jacobson et al. (2020) provide the only study exploring the causal impact of (postponing) either inductions or c-sections but they do not identify the effects of intervening vs. not intervening at all, nor do they allow for interaction effects. They find small adverse effects on neonatal health. Despite this lack of scientific evidence, non-medically indicated induction is less restricted by medical guidelines than elective surgical procedures (DGGG, 2020b,a). The new findings contradict the marginalization of induction relative to c-section.

This study also benchmarks its multi-treatment estimates against the causal literature, thereby putting 1) the reliability of the single-treatment identification on debate, and 2) the value-added of the more involved multi-treatment identification into perspective. This is especially useful for the relatively broader evidence on c-section effects, like Card et al. (2018), Costa-Ramón et al. (2018), Costa-Ramón et al. (2019), Halla et al. (2020), and Jachetta (2016). Instrumented single-treatment evidence is found to deviate substantially from multi-treatment findings. Besides, single-treatment estimates differ a lot in sign, size, and significance depending on the instrument used. Therefore, with interrelated birth interventions, a multi-treatment model yields more reliable results.

This study meets key interests of public policy. First, awareness that hospital demand for intervention aggravates rather than alleviates capacity constraints is crucial to prevent snow-balling effects generating even more birth interventions for non-medical reasons (Allen et al., 2006; Bonsack et al., 2014). Second, inferring some implications of inductions for subsequent fertility is likewise of principal interest: The suspected long-term effects of perineal damage counteract costly fertility incentives (parental leave policies, child allowances, etc.) and, most likely so, when the first child is born (Bruns, 2017).[6]

The remainder of the paper is structured as follows. section 3.2 describes the data. section 3.3 details the identification approach. section 3.4 presents estimates for the health effects of inductions while section 3.5 turns on a hospital's staff capacity effects. section 3.6 concludes by discussing the policy implications of my analysis.

3.2 Data

To analyze the effects of induced labor on patients' health and hospital staff constraints, this study uses nationwide mother-child level hospital records collected and cross-validated by the IQTIG institute.[7] The focus lies on the records from 2015 through 2016, for each

[6] On average, first births last 6.7 hours (and second births only 4.6 hours), which makes them more susceptible to intervention: all major birth interventions are way more common among first births (Table 3.1).

[7] The *Institut für Qualitätssicherung und Transparenz im Gesundheitswesen* is an independent scientific research institute with a legal mandate from the German Federal Ministry of Health to evaluate hospital

of which the pregnancy, the entire hospital stay, and the course of delivery is documented meticulously.

First, three binary treatments represent the main birth intervention types. *Induced labor*, the intervention of principal interest, is expressed as a pooled indicator showing if any form of induction has been conducted or not. It adopts the clinical definition of induced labor, which includes, most prominently, mechanical rupture of membranes and hormonal labor stimulation by medication, but excludes minor interference like cervical ripening (see Mishanina et al., 2014, for details on induction methods like Oxytocin dose or membrane sweep).[8] Second, *Non-emergency C-section* comprises all but emergency c-sections. On the one hand, this definition does not rely on possibly strategic hospital labeling of c-sections as planned vs. spontaneous (Card et al., 2018). On the other hand, excluding emergency c-sections allows focusing on c-sections with medical scope as a treatment.[9] Third, a binary indicator captures vaginal operations in a wider sense. It combines classical vaginally operative birth assistance, i.e., by forceps, vacuum, or spatula, with episiotomy, which is a surgical cut to prevent perineal damage by spontaneous tearing. Interestingly, vaginal operations are more common (32%) than inductions (28%) or c-sections (26%) in the main analysis sample (Table 3.1, column (4)). Notably, despite targeting the impact of non-medically indicated interventions, none of the treatment indicators relies on reported indications possibly manipulated to justify intervention framing low-risk births as pathological (Jürges and Köberlein, 2015; Bradford et al., 2007; Kolip et al., 2012; Feige, 2008).

The dependent variables are a binary indicators for maternal health, namely the incidence of 1) high-degree perineal tearing, as well as ordinal measures of 2) the APGAR score five minutes after birth[10], 3) the hours of labor duration, and 4) the days of the (maternal and neonatal) postnatal hospital stay. *Perineal Tearing (III/IV)* encompasses also wound hematomas, reflecting either severe perineal tearing or episiotomy itself. As episiotomy is likely not randomly assigned, it is included in the treatment *Vaginal Operations.* Impor-

care quality. Independent of public or private sponsorship, all officially registered hospitals are obliged to report their data for external validation and evaluation. In Germany as of 2010, only 2% of births took place outside a hospital (Kolip et al., 2012). It is a legal requirement to cite the data as follows. “Es wurden Daten aus Qualitätssicherungsverfahren gemäß §136 SGB V des Gemeinsamen Bundesausschusses verwendet.”

[8] Henceforth, *unassisted* labor is defined as spontaneous labor, maybe augmented or slowed down as the delivery proceeds. Likewise, *unassisted* birth precludes any of the three main birth interventions.

[9] As emergency c-sections refer to mortal danger, they are much harder to recode strategically, first and foremost due to stricter reporting requirements and a different workflow. Originally, emergency c-section was targeted as an outcome of induced labor but there was too little variation.

[10] The APGAR score (0-10) increases in healthy skin color and correct functioning of lungs, heart, muscles, and reflexes (Card et al., 2018).

tantly, even though a mother might select into episiotomy while another mother would bear perineal tearing instead, the wound hematoma signals the severe course of labor for either one.

Table 3.1: Characteristics of Births by Non-Missingness, Preconditions, and Birth Order

	1st & 2nd births	**with non-missing central variables**					
		1st & 2nd births		zero-precondition 1st births		zero-precondition 2nd births	
		= 1	Δ	= 1	Δ	= 1	Δ
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
maternal characteristics							
german (y/n)	0.80	0.81	-0.06***	0.81	0.00***	0.78	0.04***
single (y/n)	0.11	0.11	-0.07***	0.11	-0.01***	0.08	0.03***
low socioeconomic status (y/n)	0.80	0.81	-0.10***	0.80	0.01***	0.85	-0.05***
age	30.28	30.26	1.01***	28.16	3.08***	29.61	0.78***
bmi	24.72	24.72	0.15**	24.10	0.90***	24.43	0.33***
pre-pregnancy weight (kg)	68.78	68.79	-0.11	67.06	2.51***	68.02	0.89***
gestational age (#days)	275.17	275.27	-5.62***	279.81	-6.75***	279.24	-4.74***
prenatal care (#doctor visits)	11.14	11.13	0.18***	11.53	-0.57***	11.15	0.00
pre-care start >12th week (y/n)	0.07	0.07	0.00	0.08	-0.01***	0.08	-0.01***
neonatal characteristics							
birth weight (g)	3328.55	3332.10	-189.89***	3415.72	-126.43***	3527.25	-231.53***
hospital characteristics							
emerg. cs time >20 min (y/n)	0.01	0.01	0.01***	0.01	0.00***	0.02	0.00***
emerg. cs time <3 min (y/n)	0.00	0.00	0.00*	0.00	0.00*	0.00	0.00*
health outcomes							
emergency c-section (y/n)	0.01	0.01	0.01***	0.01	0.00	0.01	0.01***
perineal tearing (III/IV) (y/n)	0.01	0.01	0.00***	0.02	-0.01***	0.01	0.01***
APGAR score (5 min.)	9.6	9.6		9.7		9.8	
hospital capacity outcomes							
labor duration (#hours)	4.85	4.91	-3.47***	6.69	-2.68***	4.57	0.33***
maternal postnatal stay (#days)	3.45	3.44	0.57***	3.36	0.14***	2.66	0.93***
neonatal postnatal stay (#days)	3.12	3.12	-0.27***	3.19	-0.11***	2.59	0.62***
treatments							
induced labor (y/n)	0.22	0.23	-0.03***	0.28	-0.07***	0.20	0.03***
non-emerg. c-section (y/n)	0.31	0.34	0.32***	0.26	0.13***	0.08	0.32***
vaginal operations (y/n)	0.20	0.21	-0.01***	0.32	-0.16***	0.11	0.11***
IV staff capacity							
non-working day due date (y/n)	0.33	0.33	0.00	0.33	0.00	0.34	0.00**
PROM 8pm-4am (y/n)	0.12	0.13	-0.02***	0.15	-0.03***	0.10	0.04***
midwife shortage at arrival [0,1]	0.57	0.57	0.29***	0.58	-0.01***	0.58	0.00
IV obstetricians' preferences							
preference induced labor [0,1]	0.23	0.23	0.00	0.23	0.00***	0.23	0.00***
preference non-emerg. cs [0,1]	0.34	0.37	0.03***	0.34	0.05***	0.23	0.16***
preference vaginal operation [0,1]	0.20	0.20	-0.01***	0.20	-0.01***	0.20	0.00**
N	1,076,763	561,572		177,215		81,620	
N obstetricians' preferences	412,228	206,199		66,916		27,457	

Notes: *p<0.1; **p<0.05; ***p<0.01. IQTIG birth records for Germany 2015-2016. (Differences in) means for the central analysis variables based on all births, births restricted to non-missing central variables, and 1st and 2nd births without pregnancy or birth risks ante partum. See Table 3.A.2 and Table 3.A.3 for details on sample and variable construction. Binary indicators *yes/no* abbreviated as *y/n*. *PROM* refers to *prelabor rupture of membranes*. *emerg. cs* is short for *emergency c-section*. For the APGAR score means but no differences are available.

Given various stages of labor are observed, *Labor duration* (in hours) can be computed in terms of the staffing-relevant period a mother has contractions, as opposed to the reimbursement-relevant period of pushing contractions (DHV and DGGG, 2020).[11] The duration of the postnatal hospital stay is measured for both, a mother and a neonate, by counting the days elapsed between the completion of delivery and hospital discharge.

For causal identification, a total of six instruments is created which are discussed in depth in 3.3.3. Parsimonious baseline covariates and extensions are detailed in the notes to Table 3.A.11. Complete variable specifications are given in Table 3.A.2.

To exclude interventions planned for strictly medical reasons, the sample is restricted to zero-precondition births, i.e., mothers-to-be (henceforth: mothers) without any known pregnancy or birth risks antepartum, thereby focusing on normal presentation singleton pregnancies at term. Moreover, focusing on first births minimizes heterogeneous influences from prior parity experiences. Finally, conditioning on non-missing central estimation inputs yields the main analysis sample of 177,215 observations. Table 3.A.3 depicts an overview of sample specifications. Balance Table 3.1 compares the central variables across samples, e.g., all first and second births versus the main analysis sample by non-/missingness. Despite confirmed high data quality overall, this matters. Hospitals cannot oblige mothers to provide non-obstetric information, and column (3) suggests, e.g., the socio-economic or marital status may be selectively missing. Table 3.A.5 in the appendix balances all core characteristics across strata created for endogeneity and heterogeneity checks.

3.3 Empirical Approach

This section sheds light on the institutional background determining a hospital's incentives for physician-induced birth intervention demand and sets out the empirical strategy. After introducing a simple OLS benchmark framework, an instrumental variable strategy is developed to identify the causal impact of non-medically indicated induced labor.[12]

[11] Detailed information on labor progress is one advantage of the IQTIG data compared to other natality records. For, Card et al. (2018) do not observe labor at all and proxy its duration by counting the hours from hospital admission to completed delivery.

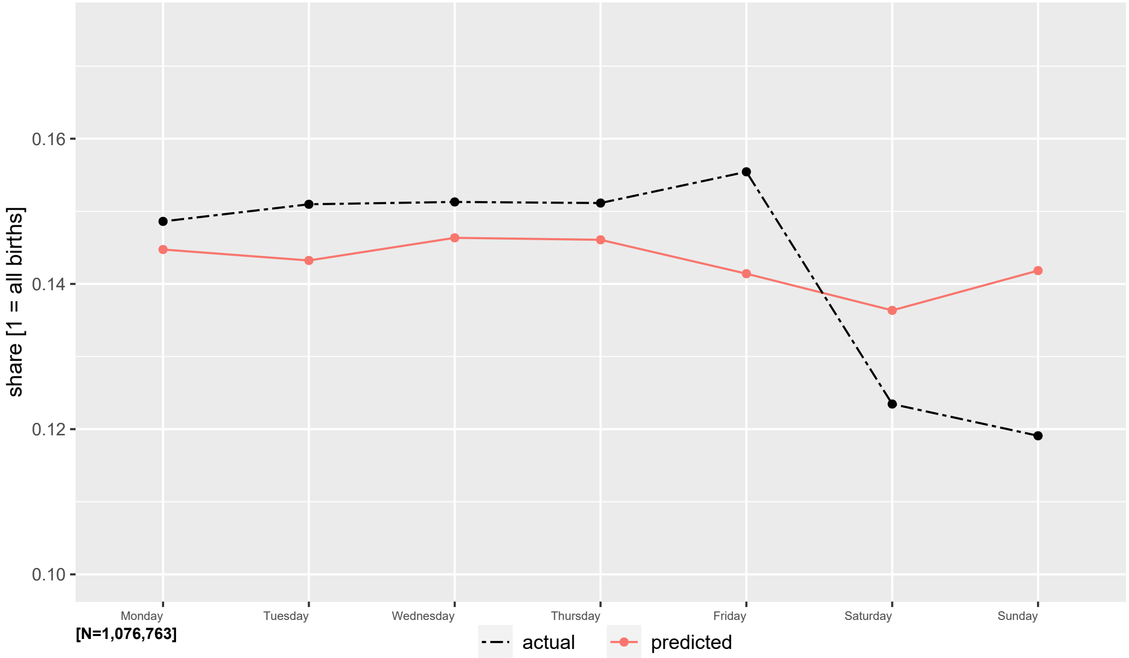
[12] Due to the highly confidential data base the empirical analysis follows a legal protocol. First, the data user commits herself to a statistical analysis plan, second, the corresponding code is run at IQTIG, and third, all output - ex-ante requested without any data insights - is released to the user after legal approval by the Gemeinsame Bundesausschuss (G-BA). Ongoing work by Gerhardts (2024) updates the analysis in response to these first findings.

3.3.1 Institutional Background & Intervention Incentives

The identification strategy exploits supply-side incentives for induction at the hospital and the obstetrician level. Inducing a woman's labor plays a key role in hospital management, first and foremost to alleviate staff shortages and to forego standby costs the hospital is not compensated for (Bruns, 2017; Feige, 2008).[13]

On behalf of an obstetrician, despite flat-rate pay, performing an induction could be appealing for many reasons.[14] The multitude of subjective incentives creates variation in obstetricians' preferences for intervention. Given decision scope from medical guidelines (Bruns, 2017) paired with variation in capacity constraints and intervention preferences, mothers are heterogeneously exposed to physician-induced demand.

Figure 3.3: Actual vs. Predicted Delivery Date Distribution Across Weekdays



Source: IQTIG German hospital records for 1st and 2nd births in 2015-2016. There are <1% of deliveries with a hospital-corrected due date (not shown). Own calculations.

In a first step, to overcome common challenges upon visualizing physician-induced

[13] Standby costs are poorly documented. Ignoring standby costs completely, the average DRG-based profits of an unassisted birth (n=100, 94 uncomplicated) are 1847-1674=173 EUR accounting for 556 EUR reimbursement-relevant staff costs (Rummel (2007)). While standby costs of just 24% (or 173 EUR) relative to 76% (or 556 EUR) would turn the profit into a loss, a more realistic standby cost estimate of up to 70% (Bruns, 2017) implies a sizable loss.

[14] Lutz and Kolip (2006) and BZgA (2005) summarize alternative forensic, demographic, economic, cultural, societal, technological, and other supply and/or demand-side incentives for inducing labor.

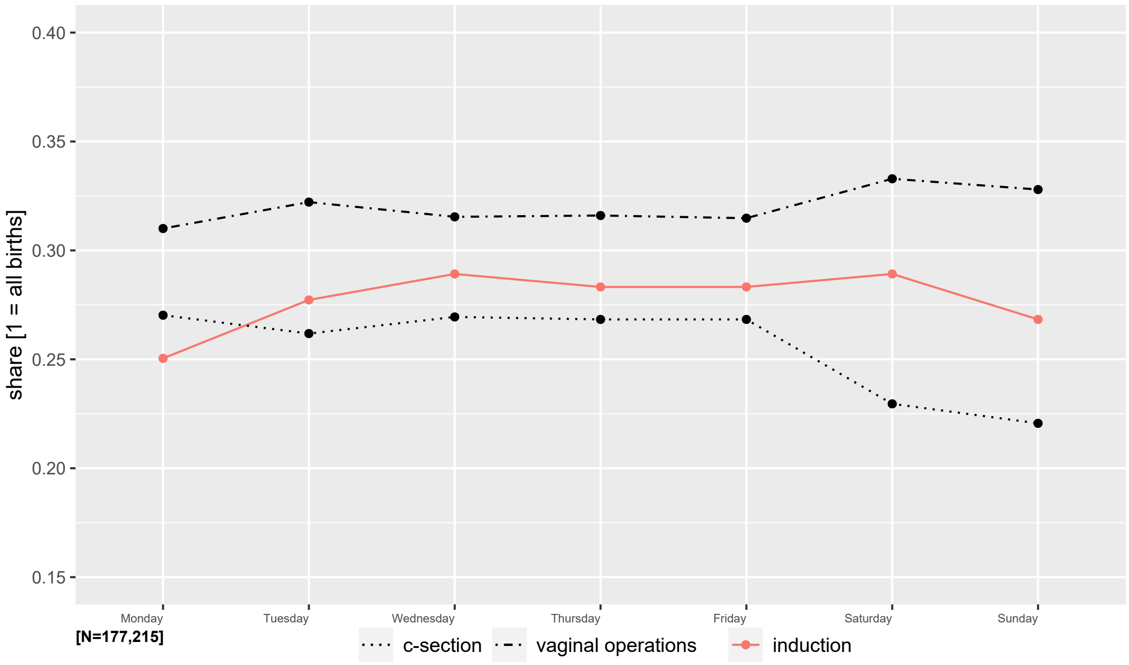
demand (Dranove and Wehner, 1994), Figure 3.3 plots due dates and completed births across weekdays. While due dates set out the biological benchmark distribution, actual (completed) deliveries should reflect potential man-made birth timing at a hospital, thereby causing diverging distributions. We see that most due dates are Wednesdays (used as the 100% benchmark) closely followed by Thursdays, while the fewest due dates are Saturdays (3% less than Fridays). The total range of fluctuation is limited to 7%. Intuitively, the non-uniform pattern could be driven by leisure time dependent menstruation cycles, etc.[15] Completed births somewhat follow the due date distribution from Monday through Thursday but running up to the weekend the patterns diverge. Most births occur on Fridays, followed by a drastic drop of 21% on Saturdays and - taking Fridays as the benchmark - a further drop of 3% on Sundays.

In a second step, to visualize work shift-specific intervention demand, Figure 3.4 and Figure 3.5 plot (the shares of) un-/assisted births across weekdays and hours of the day respectively. Induced *births*[16] are least frequent on Mondays, their share rises and stays up from Tuesdays through Saturdays before dropping down on Sundays. Induced *births* are also least common between 07 and 08 am, then their share continuously increases till peaking between 09 and 10 pm, before decreasing again. The patterns of vaginal operations seem largely mirrored by c-sections, although c-section oscillate more strongly. We see the least (most) births with vaginal operations (c-sections) on Mondays, rather stable shares throughout the week, and a distinct increase (decrease) on the weekend. Similar substitution effects can be seen across daily hours, where the fewest c-sections are performed before 06 am, then they peak already at 08 am and drop drastically starting from 03 pm. Accordingly, vaginal procedures are rarest at 08 am before overtaking c-sections in frequency at 10 am again. In short, the distribution of births following any (or a combination) of the three main intervention(s) is suggestive of some non-random timing of birth assistance.

[15] The due date prediction is normed to 40 weeks from the 1st day of the last menstruation.

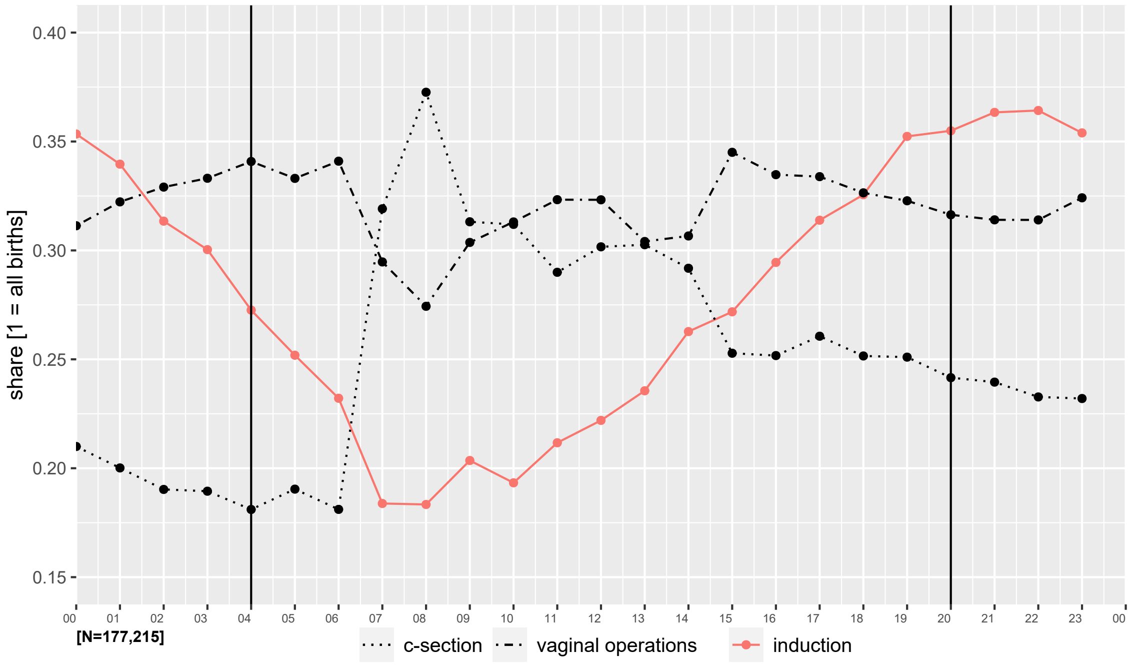
[16] To tentatively assess the timing of inductions themselves, these patterns need to be lagged by 13-17 hours (the mean interval between induction and delivery in a similar sample studied by Levine et al., 2016). For example, a 17-hour-lag (depicted in Figure Figure 3.A.3) yields an intervention pattern in line with previous literature where inductions are concentrated 1) on the early morning hours maximizing delivery likelihood during the day shift as well as 2) on Mondays through Fridays shifting delivery away from weekends (Halla et al., 2020; Costa-Ramón et al., 2018). However, inferring induction timing from birth timing is imprecise as the induction-birth interval depends on the induction method(s) and further (minor or major) interventions applied.

Figure 3.4: Relative Distribution of Assisted Births Across Weekdays



Source: IQTIG German hospital birth records for 2015-2016 restricted to the main analysis sample of zero-precondition first births (detailed in the notes to Table 3.A.11). Own calculations.

Figure 3.5: Relative Distribution of Assisted Births Across Daily Hours



Source: IQTIG German hospital birth records for 2015-2016 restricted to the main analysis sample of zero-precondition first births (detailed in the notes to Table 3.A.11). Own calculations.

3.3.2 Benchmark: Pooled OLS

As a baseline, an OLS regression with three interdependent treatments is run, thereby accounting for substitution effects and complementarities among birth interventions. Specifically, induced labor could be either substituted by a planned c-section or complemented by a spontaneous (non-emergency) c-section.[17]

Despite an ongoing debate as to whether inductions cause c-sections or not (Table 3.A.1), their joint effect has been broadly neglected. Seminal IV studies identifying c-section impacts, either ignore (Card et al., 2018), control for (Costa-Ramón et al., 2018), or drop inductions from the sample (ibid.), none of which overcomes the problem that induction is endogenous, too. Jacobson et al. (2020) distinguish unassisted, induced vaginal, and c-section deliveries, thereby mechanically mixing the impact of failed inductions with that of a c-section alone. By contrast, this new model supplements the main interventions by a single cumulative induction-plus-surgery indicator.

$$
\begin{aligned}
Y_m = \beta_0 + \beta_1 * IL_m + \beta_2 * CS_m + \beta_3 * VO_m + \beta_4 * IL_Surgery_m \\
+ \underset{1\times k}{\mathbf{x}_m^{'}} \ \underset{k\times 1}{\delta} \ + \ \underset{1\times s}{\lambda} \ \underset{s\times 1}{\mathbf{1}} + v_m
\end{aligned} \tag{3.1}
$$

where

covariates $\mathbf{x}_m \equiv \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_{k-1} \end{bmatrix}$, sets of fixed effects $\lambda \equiv \begin{bmatrix} \lambda_1 & \dots & \lambda_s \end{bmatrix}$

- Y_m: outcome of mother (or her neonate) m
- $IL_m, \ CS_m, \ \ VO_m \in \{0, 1\}$: =1 if mother m has an induction (and maybe other interventions), a non-emergency c-section (dto.), or vaginal operations (dto.) respectively; 0 else
- $IL\_Surgery_m$: =1 if mother m has an induction followed by a non-emergency c-section, vaginal operations, or both; 0 else
- details on pre-determined controls, fixed effects, and cluster-robust standard errors given below Table 3.A.11

OLS estimates for induction health impacts vary substantially in terms of size, sign, and significance (Bonsack et al., 2014; Coatesid et al., 2020; Axt-Fliedner et al., 2004). This

[17] Equation 3.3, the originally targeted model allows all two- and three-way intervention interactions. However, in practice, after *failed* vaginal operations a spontaneous (but non-emergency) c-section is medically only feasible before the fetus descends too far into the birth canal. In the main sample of 177,215 births, there are 45 doubly surgical (and 15 triple intervention) cases causing extreme multicollinearity issues. Follow-up work by Gerhardts (2024) settles on an intermediate model that dismisses rare interactions from the original specification but still manages to disentangle induction followed by c-section (9%) vs. induction followed by vaginal procedures (8% of birth modes, see Table 3.A.8).

study accounts for interrelated birth interventions and a parsimonious set of core controls inspired by Card et al. (2018). Related estimation designs by Card et al. (2018), Halla et al. (2020), Buckles and Guldi (2017), Jürges (2017), Schulkind and Shapiro (2014a), and Costa-Ramón et al. (2018) are also cautious about adding many different sets of fixed effects even though non-linear trends, seasonality of births across months, weekdays, and daily hours, as well as hospital-specific intervention effects, are well-known. Accounting for binary outcomes and possible error correlation up to the county level (due to mothers selecting into hospitals), cluster-robust standard errors are reported.

Nevertheless, the OLS benchmark regression likely yields biased estimates as interventions are non-randomly assigned. For an up-/downward bias, i.e., over-/understating the adverse health effects, inductions without medical advantages needed to be concentrated on mothers with worse/better expected health outcomes. Adverse health outcomes are consistently negatively correlated with socioeconomic status (Jeong et al., 2020). For zero-precondition first-time mothers of lower socioeconomic status more doctor visits are registered despite the belated start of prenatal care, see Table 3.A.4, which hints at curative rather than preventive appointments. In the literature, the correlation between socioeconomic status and inductions varies across countries (Carter et al., 2020).[18] Intuitively, a concentration of inductions on women with lower socioeconomic status could be rationalized by, e.g., physician-induced demand increasing in information asymmetries, thereby overstating an adverse health effect. Vice versa, an adverse health effect would be underestimated if mothers with high socioeconomic status got more high-tech medical care and thus more exposure to false positives about the fetus' well-being. Comparing the unconditional means across samples, zero-precondition first births to mothers with lower socio-economic status are equally often induced as the main sample (but more prone to c-sections), see Table 3.A.4. In line with prior literature (O'Dwyer et al., 2013; Carter et al., 2020), we further see inductions to be centered on slightly older mothers but less common among single mothers (23%). By contrast, c-sections are more frequent (around 27%) for both groups relative to the overall sample (21%). This is suggestive of (un-)observed differences likewise associated with birth outcomes even among zero-precondition first-time mothers.

[18] One key risk factor reflected in socioeconomic status is nutrition quality (Wolfe et al., 2011). Zero-precondition births exclude mothers with severe obesity and the core controls account for (even non-linear effects of) height and weight, and BMI. However, BMI is just a (noisy) proxy for dietary quality, i.e., even a slim person could have bad eating habits.

3.3.3 Instrumental Variable Estimation (IV)

To resolve self-selection into multiple, possibly combined interventions, the three major interventions and induction followed by surgical intervention are instrumented.[19] Two alternative sets of three instruments each are discussed, first in terms of exogeneity, followed by relevance.[20] A brief remark on monotonicity concludes the identification discussion.

Instrument Exogeneity & Exclusion Restriction

1. A Set of Instruments based on Obtetricians' Intervention Preferences

   This set of instruments uses an obstetrician's preferences to perform inductions, c-sections, and vaginal operations (similar to Bhuller et al. (2020) in another context). Preferences are measured via the obstetrician's average rate of performing the respective intervention in all past deliveries. The idea is that an obstetrician has both, institutional decision scope on whether to offer intervention as well as influence on a mother's consent due to physician-patient information asymmetries.

   The first requirement of the exclusion restriction is random obstetrician assignment. Addressing concerns about mothers selecting into a hospital for its intervention reputation warrants including hospital fixed effects.[21]Next, considering within-hospital randomness, restricting the sample to zero-precondition first births is important. On the one hand, zero-preconditions rule out skill-based obstetrician assignment, i.e., the matching of (unobserved) medical skills to heterogenous maternal health records. On the other hand, first-birth mothers are less likely to request a specific obstetrician based on prior experiences. However, learning about physician-specific intervention histories mothers might try to pick an obstetrician matching their preferences. Considering the organizational workload of German hospitals such selection seems unlikely but not impossible.

   Therefore, the subsample of mothers rejected by one and transferred to another hospital is analyzed because those mothers could neither choose the hospital nor the obstetrician. Table 3.A.5 shows transferred women to be more often single, older, and of higher socio-economic status, delivering relatively lower birth weight babies, and experiencing a lot more inductions or (even emergency) c-sections. Likewise, the

[19] Whether birth interventions are endogenous (and IV estimation preferable) is not tested formally because the Wu-Hausman Test/ Durbin Score does not adapt easily to this set-up. Results from a workaround exploring the regression-based approach reported after Stata's *ivregress* command (Cameron & Trivedi 2005) are available upon request.

[20] Discussing exogeneity, "as good as randomly assigned conditional on (core) covariates and fixed effects" is shortened to "randomly assigned" for the sake of simplicity.

[21] This was done only for the original model Equation 3.3 in Table 3.A.13.

subsample of mothers not presented to an obstetrician during the prenatal period is checked on. Albeit overall more similar to the main sample, these mothers are way less likely to be induced (Table 3.A.5).[22]

The second requirement of the exclusion restriction is that a given obstetrician's intervention preference may influence birth outcomes only through altered intervention likelihood. Therefore, a problematic scenario would be, e.g., a mother staying with her assigned obstetrician but refusing to collaborate with him upon learning of his preference for intervention, thereby provoking an emergency c-section. Probably more salient in this context is the gatekeeper problem (Maestas et al., 2013) meaning that obstetrician assignment could be a packaged treatment including intervention preferences but also systematic skill differences. For instance, an obstetrician's high c-section preference might result in less experience and fewer skills in handling vaginal deliveries. Reassuringly, comparing unassisted and induced labor relies on a similar skill set. Nevertheless, the sample of (otherwise low-risk) first-time mothers delivered pre-arrival to hospital could be insightful. In this sample, unassisted delivery is the default with induction (c-section) rates as low as 9% (18%, see Table 3.A.5) independent of an obstetrician's intervention preference (which is limited to influencing the type of intervention). Therefore, the estimated impact of a high induction preference should reveal other potentially influential characteristics specific to these obstetricians.[23]

2. A Set of Instruments based on Hospital Staff Capacity

Instrument: Midwife Shortages upon Maternal Arrival at a Hospital

The idea is that a mother arriving at a hospital where all midwives are busy is more likely to not get assigned a midwife at all, which in turn makes her more prone to induction.

The first requirement of the exclusion restriction is that mothers do not selectively arrive at a hospital in response to within-day minute-wise fluctuations in midwife shortages arising from ongoing deliveries there. Despite maternal midwife and hospital selection, random assignment seems plausible as a mother usually cannot observe

[22] The corresponding subsample regressions were only run for the original model Equation 3.3, see Table 3.A.12 and Table 3.A.15.

[23] Sample-specific reduced form regressions - available upon request - were only run based on Equation 3.3. Another promising subsample to test intervention preference-dependent medical skills consists of (otherwise low-risk) first-time mothers suffering pre-/eclampsia. The high blood pressure condition provokes seizures and ranges among the strongest medical indications for induction or surgical delivery. However, the sample size was <10.

current midwife shortages at a hospital, and much less so before being admitted herself. To capture unpredictable variation, midwife shortages are defined as the share of *current* deliveries without midwives.[24] The newly arriving mother herself is excluded from the shortage measure. Otherwise, if she went into labor pre-admission, her choice to bring a midwife along or not would bias the measure. Addressing concerns arising from a mother selecting into an, e.g., high-quality hospital guaranteeing a midwife to each patient upon arrival, motivates the inclusion of hospital fixed effects. Furthermore, subsample regressions for hospitals forbidding in-patient midwives rule out pre-determined mother-midwife constellations unaffected by whichever midwife shortages prevail upon hospital admission.[25]

The second requirement of the exclusion restriction is that facing a given midwife shortage may influence birth outcomes only through altered intervention likelihood. To meet this condition, the midwife shortage prevailing (not the midwife assignment itself) upon arrival should not determine whether, e.g., a mother arranges to get certain anesthesia she would not have asked for otherwise.

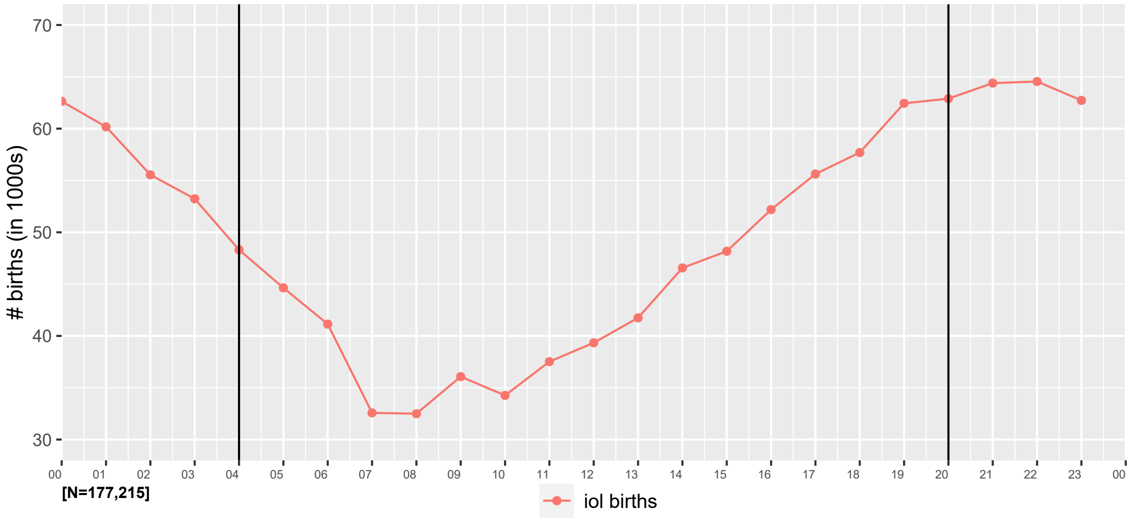
Instrument: Pre-labor Membrane Rupture during a Hospital's Night Shift

The idea is that staff shortages are relatively more acute at night making scheduling of births more attractive, especially after a membrane rupture requiring intense monitoring otherwise. Figure 3.6 plots the within-day distribution of induced *births* confirming the extent to which inductions are used to schedule births around the clock. Figure 3.7 represents a close-up of two groups, namely all births following a pre-labor membrane rupture and a subset of those that were also induced. We see that the two groups co-move to some extent, although membrane ruptures oscillate five times as strongly within a day. Both groups reach their minima around noon, five hours later than induced births overall (Figure 3.6). Thus, pre-labor membrane ruptures shape part of the induction allocation beyond obstetricians' control.

[24] Midwives are commonly assigned to several mothers simultaneously. Therefore, to capture acute and unpredictable shortages, the instrument is defined in terms of ongoing deliveries, i.e., the most care-intense periods, instead of counting midwives not assigned to a mother yet.

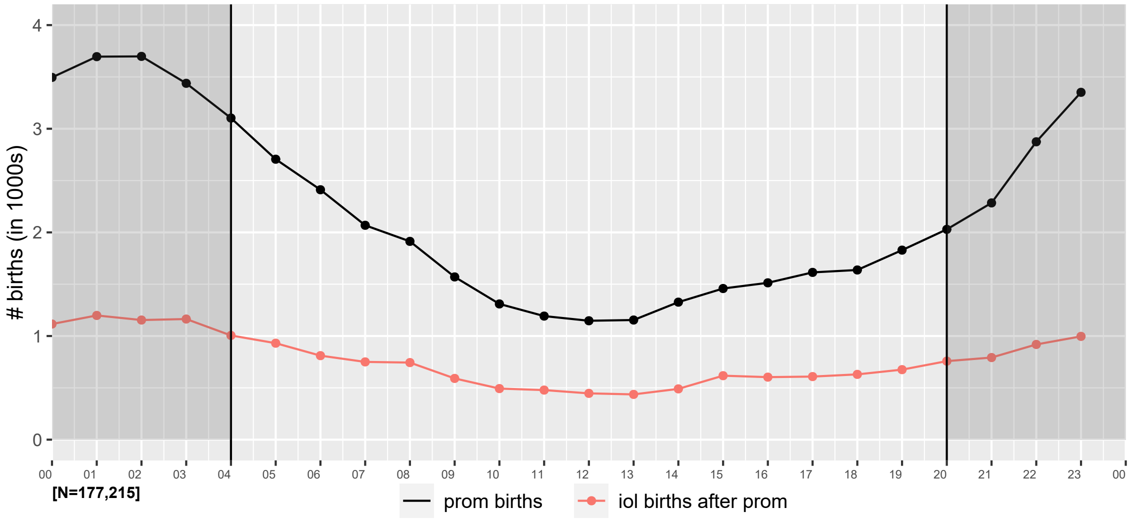
[25] This was done only for the original model Equation 3.3 in Table 3.A.13.

Figure 3.6: Induced Births Across Daily Hours



Source: IQTIG German hospital birth records for 2015-2016 restricted to the main analysis sample of zero-precondition first-births (detailed in the notes to Table 3.A.11). *iol births* refer to induced births. Own calculations.

Figure 3.7: Births with Pre-Labor Membrane Ruptures & Subsequent Inductions Across Daily Hours



Source: IQTIG German hospital birth records for 2015-2016 restricted to the main analysis sample of zero-precondition first-births (detailed in the notes to Table 3.A.11). *prom births* refer all to births following a prelabor membrane rupture, *iol births after prom* refers to a subsample of these births that are also induced. Own calculations. Please note that this figure is zoomed-in by 10 compared to Figure 3.6.

The first requirement of the exclusion restriction is that a given mother's membrane rupture falls randomly into the hospital's day or night shift. If within-day timing of membrane ruptures was randomly allocated to mothers, we would expect it to be comparable across maternal strata. Figure 3.A.4 and Figure 3.A.5 rule out influential daily working or exercising routines by plotting the daily distributions of births following membrane ruptures stratified by a mother's employment and fitness status.[26] Optimally, the timing of *membrane ruptures* (not subsequent births) should be compared as the endogenous allocation of interventions contaminates birth timing. Reassuringly, a few spikes around midday non-withstanding, even the daily distributions of births are well aligned across maternal strata.

The second requirement of the exclusion restriction is that a membrane rupture happening in either shift may impact birth outcomes only through changed intervention likelihood. That means mothers should react similarly to the rupture, e.g., by entering into the hospital as soon as possible instead of waiting for the morning to come. This seems plausible out of fear for the fetus' well-being. It is also feasible because the emergency ambulance is covered by public insurance in Germany.

Instrument: Due Date on a Non-working Day

The idea is that staff shortages are more likely on weekends and holidays, which makes scheduling labor onset for mothers due on these days relatively more appealing to a hospital.

The first requirement of the exclusion restriction is that neither parents nor physicians influence the due date's non-working day status (random assignment). In practice, the condition implies conceiving parents should not be targeting a non-working day for birth based on the due date prediction formula. Given the due date is criticized for poor precision, such a rationale seems unlikely, even if parents had non-working day preferences. However, maternal characteristics like specific working habits could influence the onset of menstruation cycles during the week. Reassuringly, across socio-economic status, no specific due date patterns emerge in Figure 3.A.2. Moreover, a gynecologist predicting the due date may not change it upon noticing a birth is due on, e.g., a Sunday. Likewise, hospitals may correct the due date prediction only for medical reasons and not to justify early-on interventions. Reassuringly, the share of mothers with a hospital-corrected due date is negligible (well below 1%).

The second requirement of the exclusion restriction is that having been assigned a non-working day due date may influence birth outcomes only through altered intervention likelihood. This condition implies that, e.g., a mother due on a Sunday may

[26] Studying membrane ruptures at $\geqslant$37 weeks of gestation implies German state-mandated maternity protection has been mitigating differential impacts of daily stress at work for about eight weeks already.

not educate herself about options of anesthesia fearing a tougher birth experience due to Sunday-specific understaffing.[27]

For conditional random instrument assignment, Figure 3.A.1 explores unconditional correlations of instruments and maternal characteristics. Many well-known patterns emerge, e.g., the intuitive overlap between instruments. Moreover, obstetricians' intervention preferences (and staffing constraints) relate to a mother's age, her bmi, whether she has a her own midwife etc., all of which strongly motivates the inclusion of core controls. Nevertheless, to put the correlations into perspective, obstetricians' preferences and staff capacity indicators are much less specific to maternal and hospital strata than the share of interventions themselves (see Table 3.A.4 and Table 3.A.5). All in all, besides controlling for the observable differences, placebo tests are warranted to assess the presence of unobservable differences possibly introducing endogeneity into the framework.

Instrument Relevance & First-Stage Results

Each set of instruments gives rise to the following system of first-stage equations.

$$\underset{t\times 1}{\mathbf{t}_m} = \underset{t\times z}{\mathbf{\Gamma}} \underset{z\times 1}{\mathbf{z}_m} + \underset{t\times k}{\mathbf{\Phi}} \underset{k\times 1}{\mathbf{x}_m} + \underset{t\times s}{\mathbf{\Lambda}} \underset{s\times 1}{\mathbf{1}} + \underset{t\times 1}{\epsilon_m} \tag{3.2}$$

Notation builds on Equation 3.1. There are $1, ..., t$ treatments, $1, ..., k-1$ covariates, and $1, ..., s$ sets of fixed effects observed for mother m, while $1, ..., z$ instruments are defined as

$$\text{either } \mathbf{z}_m \equiv \begin{bmatrix} InducedLaborPref\ (ILP_m) \\ CSectionPref\ (CSP_m) \\ VaginalOperPref\ (VOP_m) \\ ILP_m * CSP_m * VOP_m \end{bmatrix}, \text{ or } \mathbf{z}_m \equiv \begin{bmatrix} DueDateNoWorkday\ (DN_m) \\ MembraneRuptureNight\ (RN_m) \\ MidwifeShortage\ (MS_m) \\ DN_m * RN_m * MS_m \end{bmatrix}$$

- $DN_m \in \{0,1\}$: =1 if the due date is a weekend day or public holiday, 0 else
- $RN_m \in \{0,1\}$: =1 if mother m has a pre-labor membrane rupture between 8 pm to 4 am, 0 else
- $MS_m \in [0,1] = \begin{cases} 0 \text{ if \#current deliveries at that hospital} = 0 \\ \frac{\text{\#current deliveries at that hospital without a midwife}}{\text{\#current deliveries at that hospital}} \text{ else} \end{cases}$
- ILP_m, CSP_m, $VOP_m \in [0,1]$: mean prior rate of inductions, non-emergency c-sections, and vaginal operations of obstetrician treating mother m

Multi-treatment First-stage Results

[27] Anasthesia like epidurals correlate with stalled labor, emergency c-sections, and severe perineal tearing (Tammaa et al., 2007).

As classical weak instruments statistics are not applicable in this setting, underidentification is tested instead.[28] Obstetricians' intervention preferences identify Equation 3.2, i.e., underidentification is rejected by a p-value of <0.001 for all treatments (Table 3.A.6).[29] Adding to this, more intuitively than statistically, the first stage of the combined intervention *Induction + surgery* (shown in the lower panel of Table 3.2) confirms strongly significant positive correlations with all preference-based instruments.[30]

Single-treatment First-stage Results

Treating either induction or (non-emergency) c-section as the only treatment all preference-based instruments are relevant for both interventions, only induction preference is irrelevant for c-section (Table 3.2). As expected, induction (c-section) preference predicts induction (c-section) most strongly. Despite likely omitted variable bias from left-out rivaling interventions, the estimates are quite in line with intuition, e.g., they show complementarities between induction and surgical intervention preferences, as well as substitution effects between c-section and vaginal operation preferences.

Among staff capacity-based intruments, only those referring to night shift constraints predict induction (albeit with opposing signs). This is in line with the co-movement depicted in Figure 3.7.[31] All instruments but the due date's non-working day status, which

[28] Intuitively, given multiple endogenous variables, the standard first-stage F-statistic could fail as follows. Assuming a just-idenitfied model, in which one instrument is predictive of several endogenous variables, while another instrument is barely predictive for any of them. Then, in both first-stages, the F-statistic would be high, even though one of the endogenous variables would be only weakly identified. Sanderson and Windmeijer (2016) provide a cluster-robust underidentification test for multiple endogenous treatments by running the Sargan-Hansen J-Test for overidentification (Cameron and Miller, 2015) in auxiliary regressions. Conventional extensions of weak instrument statistics handle either multiple endogenous treatments, e.g., the Anderson-Rubin test (Chernozhukov et al., 2009), or cluster-robust standard errors, e.g., the Montiel-Olea-Pflueger F-statistic (Andrews and Stock, 2018) but not both.

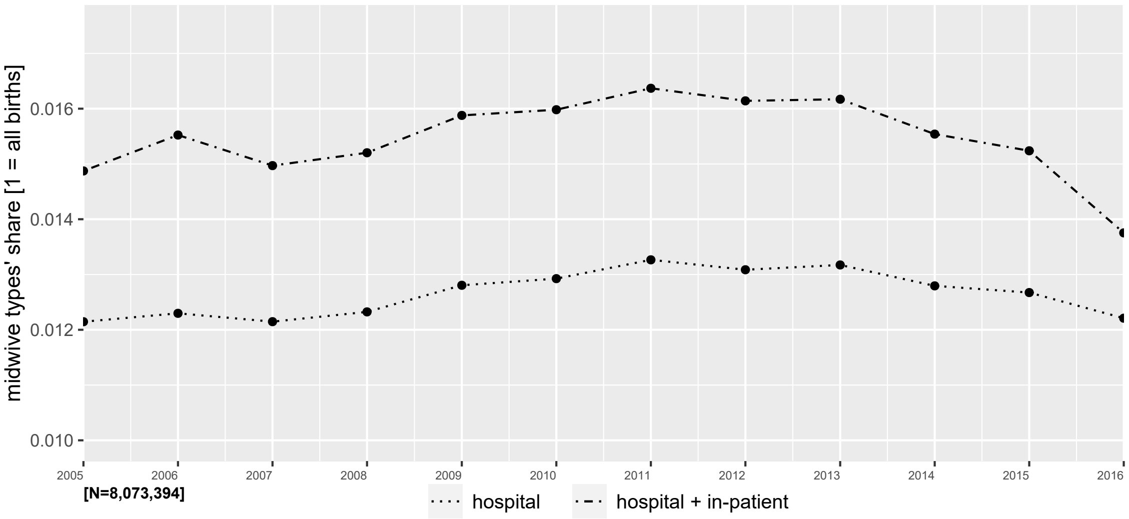
[29] By contrast, for model Equation 3.3 featuring all intervention interactions underidentification is never rejected, neither for the main nor the interacted intervention treatments (Table 3.A.8).Staff capacity-based instruments even fail to identify the parsimonious multi-treatment model (Equation 3.1).

[30] The other three first-stages - referring to each main intervention at a time - are likewise very strong and available upon request. However, the decisive criterion is the underidentfication value, not the single nor the joint significance of the first stage estimates.

[31] Absent other complications, a pre-labor membrane rupture does not imply maternal or fetal compromise. Therefore, medical guidelines state to monitor the mother closely for at least 12 hours before inducing labor (Mylonas and Friese, 2015; DGGG, 2006). However, the risk of infection increases while waiting for labor to start, and fear for the fetus' well-being probably drives down induction refusals.
Using instruments involving pre-labor membrane ruptures raises concerns about impacting a rather limited share of observations at all. 30% of zero-precondition first-time mothers experience a membrane rupture (which is way above the overall mean of 8% given by IQTIG (2017)), out of which 15% occur during the night shift (8 pm till 04 am). At the same time, given an average yearly induction rate of 28% in the sample

is never relevant anyways[32], strongly predict c-section. A significant positive correlation with midwife shortages is a common finding in the literature. BZgA (2005) and Jacobson (1993) attribute this to midwives being often more patient and more proficient in conservative obstetric skills than obstetricians. Usually, hospitals have their own midwives and/ or tolerate so-called in-patient midwives to be brought along by the mothers. However, due to severe midwife shortages, it becomes increasingly difficult to find an in-patient midwife during pregnancy (Bruns, 2017). Figure 3.8 visualizes the trend over time in midwife shortages plotting the shares of midwife types over time.

Figure 3.8: Relative Distribution of Midwive Types Across Years



Source: Destatis data downloaded from `https://www.statistischebibliothek.de` for the years 2005 through 2016. Own calculations.

Added-variable plots in the appendix (see Figure 3.A.9, Figure 3.A.10, and Figure 3.A.11) explore the residual correlation (reflected in the slope of the regression line) between a given instrument and all three main intervention types netting out maternal core characteristics. When employed in a single-treatment approach, all but one instruments capture birth intervention dynamics.

period (Schwarz et al., 2016), many mothers with zero instrument status do have their labor induced. This warrants subsample regressions exclusively for mothers with pre-labor membrane ruptures.

[32] Intuitively, the low precision of the due date counteracts instrument relevance. Still, as the single best predictor of the natural birth date, the due date represents a hospital's benchmark for treatment decisions.

Table 3.2: First-Stage Effects of Non-Medically Indicated Birth Interventions

	Staff capacity shortages					Obstetricians' preferences for			
		non-working day status of		during night shift					
Instrument	midwife shortage upon admission (1)	predicted due date (2)	actual weekday of birth (3)	prelabor membrane rupture (4)	arrival at hospital (5)	non-emergency c-section (6)	induced labor (7)	vaginal operations (8)	ind. labor x c-section x vaginal oper. (9)
Induced labor	-0.0003	-0.0022	0.0032	0.0574***	-0.2195***	0.1080***	0.2692***	0.0397**	
	(0.0029)	(0.0022)	(0.0022)	(0.0035)	(0.0028)	(0.0085)	(0.0178)	(0.0179)	
Non-emerg. c-section	0.0493***	-0.0015	-0.0397***	-0.0400***	-0.1014***	0.7308***	0.0439	-0.0675**	
	(0.0034)	(0.0022)	(0.0024)	(0.0028)	(0.0022)	(0.0089)	(0.0305)	(0.0304)	
Induction + surgery						0.2521***	0.1285***	0.1520***	1.2266***
						(0.0090)	(0.0151)	(0.0147)	(0.1693)
N	177,215	177,215	177,215	177,215	177,215	66,916	66,916	66,916	66,916

Notes: *p<0.1; **p<0.05; ***p<0.01. Linear models based on IQTIG birth records for Germany 2015-2016 restricted to the sample of zero-precondition first births (see notes to Table 3.A.11). Each row represents a birth intervention treatment to be instrumented by one (upper panel) or multiple (lower panel) instruments at a time.

3.4 Maternal & Neonatal Health Effects

This section discusses how non-medically indicated induction (possibly augmented by surgical intervention) impacts patients' immediate health. First, novel multi-treatment IV estimates are presented, then single-treatment equivalents relate to the literature. Finally, a channel-based outlook sketches possible longer-run impacts.

3.4.1 Multi-treatment IV results

Each column of Table 3.3 is dedicated to a distinct outcome. The upper four rows belong to the same specification (Equation 3.1) showing the jointly estimated impacts of all instrumented treatments, one below the other, conditional on core controls.[33] The last rows show induction estimates from separate regressions, i.e., row (5) excludes controls, and row (6) excludes the combined intervention *Induction + surgery*.[34]

Maternal Health

Following column (1) perineal laceration incidence is 6% more likely for induced (and possibly invasive) than unassisted delivery.[35] The estimated beneficial effect of induction alone is counteracted by inductions entailing a more violent course of labor and additional intervention.[36] The impact of induction alone is stable to leaving out core controls. Modeling

[33] Strictly speaking, all main effects are identified relative to unassisted delivery and deliveries suffering two-fold surgery, i.e., vaginal operations followed by c-section which occurred in the whole sample only 45 times. Assume discarding rare treatment combinations was irrelevant in terms of omitted variable bias and multicollinearity. Then, the *main* interventions' interpretation should be stable to using *Induction + surgery* instead of two- and threefold interactions.

Originally, it was intended to estimate a model using three main birth intervention treatments and all their possible interactions (Equation 3.3). However, the corresponding results indicated multicollinearity issues and underidentification. The IV (main and subsample) estimates are reported for completeness acknowledging that no causal insights arise from this evidence (Table 3.A.11, Table 3.A.12. The coefficients estimated using very weak instruments are likely more biased than their OLS analogs and inflated standard errors might render significant relationships insignificant (Angrist and Pischke, 2009).).

[34] The last row's identified effects differ, e.g., the impact of induced labor (possibly supplemented by surgical intervention) is compared to any birth mode not involving induction.

[35] Given the distribution of inductions alone vs. augmented by surgery (Table 3.A.8), the overall effect of induction in our sample can be computed as (+0.60) ppt * 0.17 = 0.102 ppt net off (-0.36) ppt * 0.28 = (-0.1008) ppt yielding (+0.0012) ppt (residual increase), relative to a sample mean of 0.02 ppt.

[36] The (only existing and imperfect) OLS benchmark (Table 3.A.11, column (2)) is based on a distinct model (Equation 3.3). and predicts a rise in perineal tearing around 0.003 ppt (15% relative to the sample mean) after induction without surgery. Notably, in this study, even the interpretation of IV and OLS effects from the same model differs. A sample of zero-precondition mothers includes inductions with debatable medical advantages, which could explain relatively more positive health effects estimated by OLS. Figure 3.A.7 and Figure 3.A.8 show the frequency of intervention indications by medical severity, which seem stable across weekdays but responsive to hours of the day.

main intervention effects only, no significant effect of induced labor (in any combination vs. unassisted delivery) emerges. Thus, disentangling the main and combined effects seems key for identification. In the literature, induction-*caused* perineal damage lacks explicit reporting (see, e.g., Jürges, 2017) preventing a direct comparison at this stage.[37] While older medical guidelines do list inductions among risk factors of severe tearing (Tammaa et al., 2007), more recent ones refer to an evidence gap about its impact (DGGG, 2020c).

Neonatal Health

Column (6) of Table 3.3 depicts a strong and stable negative intervention impact on the APGAR score five minutes (or ten minutes likewise - not shown -) post-birth, first and foremost due to induced labor (-2.2 points or 23%) but also in response to a non-emergency c-section (-0.92 points). As recapped in Table 3.A.1, seminal empirical evidence is concentrated on immediate neonatal health outcomes. Abstracting from limited comparability with single-treatment models[38], the new evidence contradicts findings by Jürges and Köberlein (2015) or Jacobson et al. (2020) and exceeds the tiny negative induction impacts suggested by Lynch et al. (2019) or Schulkind and Shapiro (2014b). Besides, the new findings speak against a positive c-section effect on the APGAR score of about 0.5 points established by Card et al. (2018) while lining up perfectly with the decrease estimated by Costa-Ramón et al. (2018). Thus, this study adds large-scale quasi-experimental evidence on adverse health effects to a highly unreconciled evidence base.

Placebo Effects

To uncover potential instrument endogeneity, this model is regressed on a battery of placebo outcomes, none of which yields significant estimates. Column (5) of Table 3.3 shows that all coefficients associated with the placebo outcome *Prenatal care starting >12th week* are insignificant. The rationale is that interventions happening at the delivery may not *lead to* events earlier throughout the pregnancy. Other placebo candidates tested are a fetus' 1) sex, and 2) innate disability, as well as a mother's 3) alcohol or cigarette abuse during pregnancy, 4) employment status, and 5) psychological or social problems.

[37] To simplify interpretation, the (relative) impact of a (non-emergency) c-section on perineal lacerations - though interpretable through a potential outcomes framework as used by Card et al. (2018) based on Abadie and Kennedy (2003) - is not discussed.

[38] Simultaneous identification within the same framework puts intervention impacts naturally into perspective to each other. By contrast, next to internal validity problems (most prominently, omitted variable bias from left-out rivaling interventions) comparing single-treatment estimates across different studies hinges on each study's external validity.

Table 3.3: Health & Capacity Multi-Treatment IV Effects of Non-Medically Indicated Birth Interventions

	patient health	hospital staff capacity			placebo	literature link
Dependent variable	perineal tearing (III/IV)	labor duration (# hours)	postnatal hospital stay (# days)		1st prenatal care >12th week	APGAR score (10 min.)
			mother	neonate		
	(1)	(2)	(3)	(4)	(5)	(6)
Induced labor	-0.3632*	-27.4370**	-0.7281	-0.3623	0.0939	-2.1522*
	(0.2035)	(11.6755)	(1.7541)	(2.1989)	(0.2621)	(1.2566)
Non-emergency c-section	-0.1512*	-17.0140***	1.8026**	1.7198*	0.0366	-0.9223*
	(0.0854)	(4.8519)	(0.7406)	(0.9354)	(0.1074)	(0.4928)
Vaginally operative procedures	-0.1391*	-10.0060**	1.2411*	1.5796*	0.0659	-0.7214
	(0.0836)	(4.6532)	(0.7103)	(0.8841)	(0.1043)	(0.4902)
Induced labor + surgery	0.5998**	40.0837***	-0.6750	-1.5735	-0.2079	2.3148
	(0.3044)	(17.2384)	(2.6272)	(3.2927)	(0.3837)	(1.7955)
Induced labor (no controls)	-0.3883*	-28.0768**	-0.5209	-0.2269	0.4036	-2.1525*
	(0.2063)	(11.7495)	(1.7291)	(2.1598)	(0.3017)	(1.2394)
Induced labor (main effects only)	0.0223	-1.6727	-1.1627***	-1.3736***	-0.0397	-0.6639***
	(0.0207)	(1.0939)	(0.2981)	(0.3351)	(0.0394)	(0.1878)
Mean (dependent variable)	0.02	6.8	3.4	3.2	0.077	9.7

Notes: *p<0.1; **p<0.05; ***p<0.01. IQTIG birth records for Germany 2015-2016. The main sample are zero-precondition first-births (N=177,215). Instrumenting by intervention preferences creates a subsample of births with non-missing obstetrician id (N=66,916). Sample and variable creation detailed in Table 3.A.2 and Table 3.A.3. The model builds on Equation 3.1, thereby using four endogenous treatments simultaneously - the main intervention indicators plus a binary indicator that is one for any induction followed by either c-section or vaginal operations or both (*Induced labor + surgery*). Jointly instrumented through obstetricians' intervention preferences and their triple interaction, the estimates of all four treatments are reported in rows (1)-(4). Besides, for induced labor, coefficients estimated by two other regressions (without controls in row (5), and excluding *Induced labor + surgery* in row (6)) are reported. Apart from row (5), regressions include as core controls the year of delivery, a mother's age, her region of origin (7 categories), her socioeconomic status (6 categories), and her single status (yes/no), where categorical variables enter as sets of binary indicators. Moreover, continuous measures are created for maternal height (as cubic), maternal weight at the beginning of the pregnancy (as cubic), and maternal BMI. Each column corresponds to a distinct outcome. Robust standard errors clustered by 3-digit zip codes of maternal residence. Means are available for the full sample of zero-precondition first-births.

3.4.2 Single-treatment IV results

Table 3.4 reports single-treatment IV estimates (conditional on core controls) for all instruments newly proposed in this study as well as some state-of-the-art instruments from the literature.[39]

Maternal Health

Using obstetricians' preferences-based instruments, having an induced (and potentially surgical) birth is estimated to increase severe perineal tearing incidence between 0.04 to 0.33 ppt relative to any birth mode not involving induction, thereby comprising not only unassisted births but even pre-labor c-sections. In line with intuition, the multi-treatment estimate (using the same instruments jointly and measuring the impact of induced vs. unassisted delivery, both of which are challenging to the perineum) is much smaller. While the impact of c-sections is of secondary interest in the context of high-degree tearing, the much smaller range (-0.05 to 0.02 ppt) somewhat puts the estimates' stability into perspective. Finally, using staff capacity-related instruments or a simple OLS model, no significant impacts are found.

Neonatal Health

Using three of five instruments at a time, induced labor predicts a significant decrease in the APGAR score (0.14-2.39 points) compared to not inducing labor. Larger impacts result from obstetricians' preferences-based instruments, the upper bound of which lines up with the preferred estimate of Table 3.3. The OLS benchmark, half the size of the IV lower bound, is significantly negative, too. When predicting the APGAR score by c-section as the only treatment the pattern is less stable, i.e., two staff capacity instruments predict a positive impact (up to 0.33 points), one capacity and one preference-based instrument suggest negative impacts of the same size, and finally, one capacity and one preference-based instrument fail to detect significant impacts at all. The naive OLS model suggests an impact overall similar to that of induction.

Placebo Tests

Interventions are interrelated, thereby responding (more or less strongly) to the same instruments. This easily turns the left-out interventions (among other candidates) into omitted variables (or bad controls, see, e.g., Costa-Ramón et al. (2018) controlling for (and stratifying by) induction upon targeting c-section effects) biasing the single-intervention IV model. Depending on which placebo outcome is chosen different drivers of endogeneity can be detected.

[39] From Table 3.2, we know that out of six proposed instruments, five identify either induction or c-section and four identify either one at a time.

Table 3.4: Health & Capacity Single-Treatment Effects of Non-Medically Indicated Induction vs. C-section

	maternal health	hospital staff capacity			placebo	literature link	
Dependent variable	perineal tearing (III/IV)	labor duration (# hours)	postnatal hospital stay (# days)		1st prenatal care >12th week	APGAR score (10 mins.)	
			mother	neonate			
	(1)	(2)	(3)	(4)	(5)	(6)	N
Induced labor (not instrumented)	0.0008	-1.0590**	0.1849***	0.1268***	0.0015	-0.0737***	177,215
	(0.0009)	(0.0449)	(0.0091)	(0.0123)	(0.0014)	(0.0047)	
Nightly prelabor membrane rupture	-0.0152	-0.4690	-0.6826***	-0.6857***	-0.0737**	0.0477	177,215
	(0.0180)	(0.7056)	(0.1800)	(0.2222)	(0.0307)	(0.0881)	
Arrival at hospital during night shift	-0.0045	-4.5616***	1.0108***	0.8340***	0.0183***	-0.1393***	177,215
	(0.0037)	(0.1487)	(0.0415)	(0.0476)	(0.0060)	(0.0175)	
Obstet.'s preference vaginal operations	0.3332*	20.0184	11.6739**	13.4012**	0.1361	-1.6002	66,916
	(0.1933)	(13.4764)	(5.6347)	(6.4334)	(0.2568)	(1.1091)	
Obstet.'s preference for c-section	0.1564***	-36.9029***	9.6169***	7.3097***	-0.1771***	-2.3933***	66,916
	(0.0269)	(3.0332)	(0.8717)	(0.7897)	(0.0453)	(0.2520)	
Obstet.'s preference for inductions	0.0361*	-1.6921	-0.4765*	-0.6667**	-0.0366	-0.7348***	66,916
	(0.0203)	(1.2240)	(0.2868)	(0.2981)	(0.0364)	(0.1799)	
Non-emergency c-section (not instrumented)	-0.0333***	-4.9841***	1.2523***	1.0687***	-0.0025*	-0.0946***	177,215
	(0.0008)	(0.0897)	(0.0153)	(0.0181)	(0.0015)	(0.0061)	
Nightly prelabor membrane rupture	0.0218	0.6731	0.9797***	0.9841***	0.1058**	-0.0685	177,215
	(0.0259)	(1.0245)	(0.2361)	(0.2993)	(0.0444)	(0.1260)	
Arrival at hospital during night shift	-0.0097	-9.8733***	2.1879***	1.8051***	0.0396***	-0.3010***	177,215
	(0.0079)	(0.3050)	(0.0864)	(0.1017)	(0.0129)	(0.0383)	
Obstet.'s preference vaginal operations	-0.1959*	-11.7677**	-6.8624*	-7.8778*	-0.0800	0.9553	66,916
	(0.1134)	(5.0723)	(3.8674)	(4.1733)	(0.1418)	(0.8008)	
Obstet.'s preference for c-section	0.0231***	-5.4556***	1.4217***	1.0806***	-0.0262***	-0.3536***	66,916
	(0.0034)	(0.2148)	(0.0535)	(0.0697)	(0.0065)	(0.0298)	
Midwife shortage upon arrival	-0.0489***	-20.8249***	4.5028***	4.6815***	-0.0152	0.3324***	177,215
	(0.0157)	(1.1067)	(0.3331)	(0.3758)	(0.0298)	(0.1132)	
Weekday of delivery non-working day	-0.0190	-10.7459***	1.1859***	1.3024***	-0.0229	0.2304**	177,215
	(0.0196)	(0.7736)	(0.1834)	(0.2386)	(0.0332)	(0.1017)	
Mean (dependent variable)	0.02	6.8	3.4	3.2	0.077	9.7	

Notes: *p<0.1; **p<0.05; ***p<0.01. models based on IQTIG birth records for Germany 2015-2016. The main sample are zero-precondition first-births (N=177,215). Instrumenting by intervention preferences creates a subsample of births with non-missing obstetrician id. Sample and variable creation detailed in Table 3.A.2 and Table 3.A.3. The just-identified specification reduces Equation 3.1 to a single endogenous treatment (cond. on core controls). Each column corresponds to a distinct outcome. Rows (1) and (7) show OLS estimates of induction and c-section, respectively. Rows (2)-(6) and (8)-(13) report IV estimates of, e.g., induced labor instrumented by a nightly prelabor membrane rupture (row (3)). Robust standard errors clustered by 3-digit zip codes of maternal residence. Means are available for the full sample of zero-precondition first-births.

Therefore, it is not surprising that instruments yield significant impacts of induction (or c-section) on belated prenatal care (and other placebo outcomes). Taking this evidence and the reasons laid out in section 3.3 together, single-treatment IV models should be interpreted with caution.

3.4.3 Channels of Longer-run Induction Impacts

Severe Perineal Tearing & Future Surgery

A growing body of observational literature raises concerns about *Perineal tearing (III/IV)* impairing maternal future health and fertility outcomes. Especially debated is the strong association with future surgery, most prominently due to 1) recurrence of tearing (Priddis et al., 2013; Woolner et al., 2019), 2) subsequent c-section on request (O'Donovan and O'Donovan, 2018; Størksen et al., 2015; Ryding et al., 2016; Smarandache et al., 2016; Garthus-Niegel et al., 2014; Pang et al., 2008; Tschudin et al., 2009; Woolner et al., 2019), as well as (3) avoidance of future pregnancies (Priddis et al., 2013) or even infertility (Jolly et al., 1999; Gottvall and Waldenström, 2002), which is especially critical as severe tearing is centered on first-time mothers (DGGG, 2020c). However, due to the lack of quasi-experimental multi-treatment evidence, this channel forbids causal longer-run inference of induction impacts.

Induction & Subsequent C-section: A Thought Experiment

But to which extent are inductions burdening the health care system by causing c-sections?[40] For a back-of-the-envelope quantification within the scope of this study, let's assume the share of mothers with induction-debited c-sections to be 7%.[41] On the one hand, induced

[40] Rummel (2007) computes hospital profits (based on DRG cost-rates among n=100 mostly uncomplicated births) for a c-section (3,843 EUR [reimbursement] - 2,385 EUR [reimbursement-relevant hospital costs] = 1,458 EUR) vs. an unassisted vaginal birth (1,847 EUR - 1,674 EUR = 173 EUR). Thus, c-section-specific additional reimbursement amounts to 3,843 EUR - 1,847 EUR = 1,996 EUR.
Using register data of *BARMER GEK*, a major German public health fund in 2010, Kolip et al. (2012) find average reimbursement costs of 1,520 EUR vs. 2,680 EUR for vaginal and c-section delivery respectively implying 1,160 EUR additional reimbursement for c-section. Despite their differences, the studies agree on a sizable extra financial burden of c-sections for the health insurance system.

[41] In our sample of healthy mothers, 79% of inductions happen on non-medical grounds (Figure 3.A.7 and Figure 3.A.8). The share of elective inductions is derived from the composition of indications (i.e., summing up explicit maternal requests and incompletely specified risks) reported for inductions in our sample. While indications may be strategically coded (Jürges and Köberlein, 2015), intuitively, hospitals should rather understate elective interventions implying a lower-bound cost estimate. For simplicity, imposing the share observed for induced on all assisted zero-precondition first-births, 79% * 9% (the share of c-sections immediately following induction) = 7% of healthy mothers experience both, induction and c-section without a medical indication ex-ante. Per 1000-women, drawing on Rummel (2007), we get 1,000 * 7% * 1,400 EUR

and unassisted labor can fail alike, which breaks the direct link between c-section as an outcome of labor trials. On the other hand, the calculation is conservative in accounting only for immediately provoked c-sections, thereby excluding higher-order parity c-sections (maybe resections[42]) requested due to traumatic past induction. Expressed per 1000-women, this yields 99,500 EUR profits for the hospital while burdening the health care system with 140,000 EUR. As of 2019, across healthy first-time mothers, this implies 34 million hospital profits vs. 47.8 million losses for the public health care system.[43]

3.5 Hospital Staff Capacity Effects

This section discusses the impact of non-medically indicated induction and/ or surgical intervention on a hospital's staff capacities. Adopting a structure similar to section 3.4, the focus lies on new causal evidence from a parsimonious multi-treatment IV model, which is put into perspective by seminal single-treatment models from the literature. Some back-of-the-envelope calculations assessing the system-wide health care impact conclude.

3.5.1 Multi-treatment IV Results

Do the adverse intervention effects established in section 3.4 rebound from patients' impaired health onto hospitals' staff capacity constraints? Columns (2) to (4) of Table 3.3 based on Equation 3.1 quantify staff capacity impacts via two key measures of a hospital's monitoring workload.[44]

Labor Duration

For inductions (possibly augmented by surgery) average labor shortens by close to 1 hour.[45] The pure induction effect is stable to leaving out core controls. Using main effects alone, no significant effect of induction emerges. Surgical interventions mechanically cut labor

(hospital profit per c-section) = 99,500 EUR vs. 1,000 * 7% * 2,000 EUR (additional health care burden per c-section) = 140,000 EUR.

[42] With a share of 23.6% in 2010 already (Kolip et al., 2012), resection has been the most popular indication for c-section for more than a decade.

[43] Kolip et al. (2012) suggest 5% of first-time mothers suffer preconditions. Using 2019 as base year, this equals 341,000 zero-precondition first births among 359,000 first-time mothers from Germany. The actual sample size used in this study deviates due to a different sample period, a (stricter) definition of zero-preconditions, and some missings in non-mandatory maternal background information.

[44] Results of the originally intended model (Equation 3.3) are not discussed.

[45] Drawing on the intervention shares among zero-precondition first-time mothers (Table 3.A.8), we find some 0.28 * (-27.4) h = -7.7 hours shorter labor due to induction alone vs. a (0.09 + 0.08) * (+40.1) h = 6,8 hours prolongation, which yields (-0.9) hours (-13%) of foregone labor experienced by a representative mother.

short by 10 (vaginal operations) to 17 (c-section) hours. All in all, interventions produce overarching favorable effects for staff capacity absorbed by *Labor Duration*.[46]

Postnatal Hospital Stay

Neither sole nor combined induction impact the postnatal hospital stay significantly. Notably, the accompanying standard errors are unusually large. Not including controls, standard errors are slightly smaller and the estimated coefficients decrease (in absolute terms) by around one third. The main-effects-only model yields highly significant negative impacts that shorten the maternal (neonatal) hospital stay by 1.2 days or 35.3% (1.4 days) following induction in any combination relative to any not induced birth.

The hospital stay of mothers and neonates is prolonged after births involving c-section (ca. 1.8 days for both) or vaginal operations (1.2 and 1.6 days). Thus, surgical interventions alone drive adverse health effects mirrored in additional patient monitoring. This might explain (part of) the marginalization of inductions relative to surgical birth interventions reflected in inductions' less restricted usage on non-medical grounds.

3.5.2 Single-treatment IV results

Regressing hospital staff capacity outcomes on induction modeled as single treatments (Equation 3.3) yields mostly significant but volatile estimates across instruments.

Labor Duration

According to column (2) of Table 3.4 induction (relative to any other birth mode) is estimated to significantly shorten labor duration between -36.9 to -4.6 hours. The wide range relies on just two (out of five) instruments. Upon instrumenting by obstetricians' preferences precision is a problem. Especially for vaginal operations' preference standard errors explode across all staff capacity outcomes. The corresponding OLS estimate predicts a significant but relatively modest decrease in labor length by 1 hour, well aligned with the (conceptionally different) total impact the multi-treatment specification Equation 3.1 yields.

For all but one instrument, non-emergency c-section is estimated to significantly shorten labor between -20.8 to -5.5 hours. Centered around 10 hours, the range encompasses the -17 hour decrease found by the multi-treatment estimate for c-section (not mixed with trial of labor and compared to unassisted birth). Finally, the OLS estimate is close to the IV lower bound estimate (5 hours).

[46] The health impact of labor length is ambiguous. On the one hand, longer labor causes longer pain and exhaustion. On the other hand, shorter labor might come at the cost of severe tearing or even phenomena like precipitate deliveries. Viewing shorter labor in the light of worse tearing (section 3.4) speaks of a hastened birth experience and disadvantage for the patient.

Postnatal Hospital Stay

The estimated impact of induction on patients' postnatal hospital stay is always significant. The predicted positive impacts (accompanied by inflated standard errors) reach up to 13 extra days in the hospital. Depending on the instrument some much smaller negative effects emerge. The corresponding OLS estimates are negligible in size and the multi-treatment model (Table 3.3) finds no significant induction-related effects at all.

For c-sections, the most precise single-treatment estimates suggest 1 - 5 additional days in the hospital encompassing the multi-treatment estimate of 1.8 days shown in Table 3.3. The only negative (and very imprecise) coefficients emerge when instrumenting by vaginal operations' preference (-7 days). OLS estimates are consistently positive and close to 1 additional day in the hospital for both, mothers and neonates.

3.5.3 Back-of-the-Envelope Calculation: Compound Costs for Intervention-Related Monitoring

The DGGG (2020a) states that elective inductions' financial impact on the health sector has not yet been established. As for (elective) c-sections, Feige (2008) mentions 100 million EUR annual reimbursement burden. Based on Table 3.3, this section sketches German hospitals' system-wide *staffing capacity burden* originating from the major birth interventions performed to relieve their *staffing constraints*.

Labor Duration

Per 1000-women, non-medically indicated interventions forego ca. 8,500 hours of labor or >1.6 million staffing costs.[47] Considering 341,000 healthy first-time mothers in Germany as of 2019 (subsection 3.4.3), foregone labor saves 547 million EUR staffing costs, a hazardous misalignment of maternal and hospital interests.

Postnatal Stay

Per 1000-women-and-neonates, additional costs for a prolonged stay of ca. 1,800 days implies a financial burden of 551,000 EUR. For all zero precondition first-time deliveries, this amounts to 188 million EUR.[48]

[47] 1000 * (28% * (-27.44) h (induction) + 26% * (-17.01) h (c-section) + 32% * (-10.00) h (vaginal operations) + 17% * (+40.08) h (induction-plus-surgery)) = (-8492.2) hours. Following Rummel (2007), this implies (2/3[h] * 50 EUR + 1[h] * 40 EUR) = 622,761 EUR reimbursable costs => -2,075,871 EUR total monitoring costs [100%, incl. non-reimbursable] drawing on Bruns (2017), 1.6 M (79%) of which are presumably non-medically indicated (subsection 3.4.3).

[48] Per 1000-women, 1000 * (26% * +1.80 days (c-section) + 32% * +1.24 days (vaginal operations)) =

Adverse Health Effects & Implicit Staffing Burden

Focusing on a hospital's short-run capacity costs as reflected in the main outcomes of this study yields a computation conservative in several ways. First, using mean cost rates associated with uncomplicated births abstracts from potential adverse health outcomes requiring not just prolonged but also *more intense* monitoring like, e.g., neonatal ventilation. Second, explicitly ignoring many other cost types is bound to underestimate the full costs by far.

The German guideline system proposes hospital care procedure workflows, thereby mapping staffing obligations to adverse health conditions. Among healthy first-births, the mean APGAR score is 9.7 (sd 0.76) and the estimated decrease due to induction is 2.15. A score of <8 warrants additional testing already (GNPI, 2022). Still conservative, we assume two additional basic tests performed per mother's labor induced on non-medical grounds. Then, among 341,000 healthy first-time mothers testing due to non-medically indicated induction entails total hospital costs of 11.8 million EUR.[49] Moreover, if the results of these routine tests confirmed neonatal adaptation anomalies more involved testing and care procedures would follow (GNPI, 2022). Modest in absolute values, this exemplary induction impact channeled by *two routine tests* is already close to 12% of the annual burden attributed to avoidable c-section procedures as a whole. As soon as intensive care measures come into play, costs rise astronomically (Almond et al., 2005).

Taken together, the naive computations have shown 1) that (weakly) favorable induction effects on seminal hospital capacity measures do not rule out substantial negative impacts on staffing costs working through more subtle channels. Besides we learn, 2) how rapidly a hospital's costs diverge from the cost it is compensated for. This in turn incentivizes more intervention without medical reason fueling a snow-balling effect that, in the long run, suggests adverse impacts on hospitals and patients alike.

+864.8 days. Per 1000-neonates, 1000 * (26% * (+1.72) days (c-section) + 32% * (+1.58) days (vaginal operations)) = (+952.8) days. For both jointly, (+1817.6) days * (30 EUR + 1/2h * 50 EUR) = 99,968 reimbursable costs [=30%] => 333,227 EUR [100%] total costs (Bruns, 2017). Proxying accommodation base costs of 200 EUR/day = 363,520 EUR, total staffing + accommodation costs = 696,747 EUR, out of which 551 k EUR 79% (see subsection 3.4.3) arise from non-medically indicated interventions.

[49] Applying current cost rates (KBV, 2020) per 1000-women, we get 1000 * 28% (induction rate) * 79% (non-medically indicated) * (17 EUR for pulse oximetry + 30 EUR for an electrocardiogram)= 10,4 k EUR [30%] => 34.6 k EUR [100% incl. non-reimbursable monitoring (Bruns, 2017)] total hospital costs for testing due to inductions performed on non-medical grounds.

3.6 Discussion

Non-medically indicated labor induction is a viral topic around the world. This chapter provides novel evidence for the role of inductions performed to alleviate staff capacity constraints in German hospitals. The estimations shown here are based on a new identification approach that uses exogenous variation in obstetricians' intervention preferences, ruling out key concerns of endogeneity through a battery of placebo tests. The main results document that induced vs. unassisted labor 1) provokes severe birth canal lacerations and lower APGAR scores, which rebound on staff capacity via 2) additional examinations and monitoring.

A framework incorporating the endogenous and interrelated nature of the three major birth interventions is pioneer work in the field. To begin with, interactions isolate successful and failed inductions entailing surgical intervention. Next, regarding the marginalization of induction relative to surgical interventions, simultaneous impact identification makes it trivial to compare the effects to each other. Last, methodologically cleaner than prior literature, the framework benchmarks evidence from single-treatment identification.

Tentatively sketching *some* likely follow-up costs for the public health care system touches upon the unresolved link between inductions and subsequent (c-section) surgery. Apart from the overall health impact explored here, ongoing work by Gerhardts (2024) focuses on heterogeneous impacts across different types of mothers. If induction-related lower APGAR scores were centered on mothers with, e.g., lower socioeconomic status, this would impair neonatal health and cognitive development disproportionately. Furthermore, examining intervention effects at low-quality versus small hospitals will shed light on a disputed reason for centralizing maternity care.

Finally, discussing the (adverse) health effects of birth interventions in the light of the professional ethical ideal stated in the very first citation, goes beyond the scope of an economics study.

3.A Appendix

Table 3.A.1: Quasi-/Experimental Evidence on Non-medically Indicated Induction

Outcome	**Study authors**	**Design**	**Health Impact** positive (+)/ neutral (=)/ negative (-)
Maternal			
Labor progress	Buckles et al (2017)	IV elective delivery policy (n=410,459)	(−) precipitious labor more likely (4.6 x) [$<$ 39 weeks]
	Jacobson et al (2020)	Reduc.Form holiday effect (n=4,599)	(=) labor complications
Blood loss	Saccone et al (2015)	SRMA of 5 ARRIVE trials (n=844)	(+) -58 mL
Surgical delivery: C-section (CS) & vaginal operations	Buckles et al (2017)	IV elective delivery policy (n=410,459)	(=) CS [$<$ 39 weeks]
	Jürges (2018)	DID parental leave policy (n=565,000)	(=) emergency CS
	Alfirevic et al (2009)	SRMA of 61 ARRIVE trials (n=12,819)	(+/−) less failed labor (8.4%:53.8%)/ more epidurals
	Wood et al (2014)	SRMA of 31 ARRIVE trials (n=12,166)	(+) fewer CS [w/o membrane rupture]
	Saccone et al (2015)	SRMA of 5 ARRIVE trials (n=844)	(=) CS
	Mishanina et al (2014)	SRMA of 157 ARRIVE trials (31,085)	(+) CS less likely (-12%) [$\geq$ 39 weeks]
	Sanchez et al (2003)	SRMA of 16 ARRIVE trials (n=6,588)	(+) fewer CS (20.1%:22.0%) [at 41 weeks]
	Middleton et al (2018)	SRMA of 27 ARRIVE trials (n=11,738)	(+/−) fewer CS/ more vaginal operations [$\geq$ 39 weeks]
	Gülmezoglu et al (2012)	SRMA of 21 ARRIVE trials (n=8,749)	(+) fewer CS [$\geq$ 39 weeks]
	Caughey et al (2009)	SRMA of 9 ARRIVE trials (n=6,138)	(+) fewer CS [at 41 weeks]
	Dare et al (2018)	SRMA of 12 trials (n=6,814)	(=) CS [w membrane rupture]
	Miller et al (2015)	ARRIVE trial (n=162)	(=) CS
	Wennerholm et al (2009)	SRMA of 13 ARRIVE trials (n=5,920)	(+) fewer CS [$\geq$ 41 weeks]
	Sotiriadis et al (2019)	SRMA of 5 ARRIVE trials (n=7,261)	(+) fewer CS
	Grobman et al (2018)	ARRIVE trial (n=6,106)	(+) fewer CS (18.6%:22.2%)
Infection	Dare et al (2018)	SRMA of 4 trials (n=445)	(=) uterine [w membrane rupture]
	Dare et al (2018)	SRMA of 9 trials (n=6,611)	(=) placental [w membrane rupture]
Neonatal			
Infection	Dare et al (2018)	SRMA of 12 trials (n=6,406)	(=) [w membrane rupture]
Birth injury	Buckles et al (2017)	IV elective delivery policy (n=410,459)	(−) more likely (8 x) [$<$ 39 weeks]
APGAR score (5 minutes postpartum)	Saccone et al (2015)	SRMA of 5 ARRIVE trials (n=844)	(=)
	Sanchez et al (2003)	SRMA of 16 ARRIVE trials (n=6,588)	(=) [at 41 weeks]
	Middleton et al (2018)	SRMA of 16 ARRIVE trials (n=9,047)	(+) fewer APGAR $<$7 [$\geq$ 39 weeks]
	Lynch et al (2019)	RDD Baby Bonus (n=1,862)	(−)
	Jürges (2018)	DID parental leave policy (n=565,000)	(=)
	Jacobson et al (2020)	Reduc.Form holiday effect (n=4,599)	(=)
	Schulkind et al (2014)	NatEx tax benefit (n=44,389)	(−) fewer normal APGAR scores at antedated birth
Birth weight	Buckles et al (2017)	IV elective delivery policy (n=410,459)	(−) -251 g [$<$ 39 weeks]
	Lynch et al (2019)	RDD Baby Bonus (n=1,862)	(−) not postponing birth, $<$2500 g more likely
	Jürges (2018)	DID parental leave policy (n=565,000)	(=)
	Sotiriadis et al (2019)	SRMA of 5 ARRIVE trials (n=7,261)	(−) -81 g
	Grobman et al (2018)	ARRIVE trial (n=6,106)	(−) lower median birth weight
	Gans et al (2008)	NatEx: new Baby Bonus (n=1,040)	(−) -75 g
	Jacobson et al (2020)	Reduc.Form holiday effect (n=4,599)	(−) -2 g
	Hussain et al (2011)	SRMA of 14 ARRIVE trials (n=6,597)	(+) fewer births $\geq$ 4000 g [at 41 weeks]
Respiratory issues	Buckles et al (2017)	IV elective delivery policy (n=410,459)	(−) 4 x more likely [$<$ 39 weeks]
	Lynch et al (2019)	RDD Baby Bonus (n=1,862)	(−) not postponing birth, normal breathing later
	Sotiriadis et al (2019)	SRMA of 5 ARRIVE trials (n=7,261)	(+)
	Jacobson et al (2020)	Reduc.Form holiday effect (n=4,599)	(=)
Mortality	Hussain et al (2011)	SRMA of 14 ARRIVE trials (n=6,597)	(+/ =) fewer deaths / equal stillbirths [at 41 weeks]
	Wennerholm et al (2009)	SRMA of 11 ARRIVE trials (n=5,920)	(+) fewer deaths [at 41 weeks]
	Saccone et al (2015)	SRMA of 5 ARRIVE trials (n=844)	(=) deaths
	Sanchez et al (2003)	SRMA of 16 ARRIVE trials (n=6,588)	(=) [at 41 weeks]
	Middleton et al (2018)	SRMA of 20 ARRIVE trials (n=9,960)	(+) deaths (2:16)/ stillbirths (1:10) [$\geq$ 39 weeks]
	Gülmezoglu et al (2012)	SRMA of 17 ARRIVE trials (n=7,407)	(+) deaths (1:13) [$\geq$ 39 weeks]
	Jürges (2018)	DID parental leave policy (n=565,000)	(=) death in first 7 days
Hospital/ intensive care visits	Lynch et al (2019)	RDD Baby Bonus (n=1,862)	(−) more visits for respiration
	Saccone et al (2015)	SRMA of 5 ARRIVE trials (n=844)	(=) #intensive care visits
	Sanchez et al (2003)	SRMA of 16 ARRIVE trials (n=6,588)	(=) #intensive care visits
	Middleton et al (2018)	SRMA of 13 ARRIVE trials (n=8,531)	(+) fewer intensive care visits [$\geq$39 weeks]
	Gülmezoglu et al (2012)	SRMA of 10 ARRIVE trials (n=6,161)	(=) #intensive care visits [$\geq$39 weeks]
	Dare et al (2018)	SRMA of 12 trials (n=6,814)	(+) fewer intensive care visits [w membrane rupture]
	Jürges (2018)	DID parental leave policy (n=565,000)	(=) #hospital visits
	Jacobson et al (2020)	Reduc.Form holiday effect (n=4,599)	(=) #intensive care visits

Notes: SRMA = systematic review & meta-analysis. Defaults: ARRIVE trials (randomizing induction vs. awaiting labor onset); natural experiments (postponing scheduled interventions) on singleton 39 weeks gestations. Deviations: marked, e.g., [$<$39 weeks] for preterm induction.

Table 3.A.2: Overview Variable Specification

Variable	**Specification**	**Function**
Maternal characteristics		
region of origin	Germany=0; Middle/Northern Europe, North America=1; Mediterranean Countries=2; Eastern Europe=3; Middle East (incl. North Africa); 5=Asian (excl. 4); 9=other	core fixed effects
residence (state-level)	[1,16]	merge w holiday data
residence (3-digit zip code level)		clustering
single status	Binary indicator =1 if mother is single; 0 else	core controls
socio-economic status	Housewife=1; apprenticeship/college enrolment=2; un-/semiskilled workers=3; lower civil servants, employees w executing responsibilities, self-employed w small business=5; (at least) intermediate civil servants, employees w (at least) extensive responsibilities, self-employed w (at least) medium business, master, site foreman, overseer=6; unknown=9	core fixed effects
socio-economic status low	Binary indicator = 1 if socio-economic status=4; 0 if status=6	strata
employed	Binary indicator = 1 if mother is employed: 0 else	core controls
age	Age in years [18,35]	core fixed effects
older age	Binary indicator = 1 if age >25; 0 else	strata
weight^3	Pre-pregnancy weight (kg) as cubic	core controls
height^3	Height (cm) as cubic	core controls
bmi	Pre-pregnancy weight / (height/100)^2	core controls
bmi ⩾ 90%ile	Binary indicator = 1 if BMI in 90%ile (all births); 0 else	strata
gestational age	#Days	miscellaneous controls
prenatal care	#Doctor visits	miscellaneous controls
prenatal care begin >12th week	Binary indicator = 1 if 1st prenatal visit >12th week of pregnancy	miscellaneous controls, placebo outcome
met obstetrician during pregnancy	Binary indicator = 1 if mother met obstetrician earlier in pregnancy	strata
hospital stay during pregnancy	Binary indicator = 1 if hospital stay earlier during pregnancy; 0 else	miscellaneous controls, placebo outcome
admitted after transfer	Binary indicator = 1 if transfer and receiving hospital id; 0 else	strata
year of completed delivery	Factor variable [2004;2019]	core fixed effects
month of completed delivery	Factor variable	additional fixed effects
weekday of completed delivery	Factor variable	additional fixed effects
hour of completed delivery	Factor variable	additional fixed effects
eclampsia	Binary variable = 1 if a mother has eclampsia; 0 else	strata
zero-precondition birth	Binary variable = 1 if non-risky pregnancy (gynecologist's label), single fetus, correct presentation, ⩾ 37 gestation weeks, no prior uterine scar, no eclampsia, no growth restriction, age 18-35, BMI <90%tile, and <20 prenatal visits ; 0 else	strata
Neonatal characteristics		
birth order	Computed as the #previous (live + still) births +1	strata
birth weight	Measured in (g)	miscellaneous controls
body measures low	Binary indicator = 1 if weight <2.5 kg, length <45, or head circumf. <32 cm	
Hospital characteristics		
hospital id	Hospital identifier	additional fixed effects
emergency c-section time >20 min	Binary indicator = 1 if condition holds; 0 else	hospital controls
emergency c-section time <3 min	Binary indicator = 1 if condition holds; 0 else	hospital controls
hospital quality low	Binary indicator = 1 if emergency c-section time >20 min | <3 min	
hospital small	Binary indicator = 1 if hospital-year specific #obstetricians <median #obstetricians p.a.; 0 else	
hospital w/o in-patient midwives	Binary indicator = 1 if #deliveries w in-patient midwives >0; 0 else	strata
Health outcomes		
emergency c-section	Binary indicator = 1 if mother needs an emergency c-section; 0 else	dependent variable
perineal tearing (III/IV)	Binary indicator = 1 if mother suffers high-level perineal damage; 0 else	dependent variable
APGAR score (5 min. postbirth)	ordinal [0-10] neonatal fitness measure (10 being top score)	dependent variable
Hospital staff capacity outcomes		
labor duration	#Hours	dependent variable
pushing contractions	#Minutes	descriptives
maternal postnatal hospital stay	#Days from completed delivery till discharge	dependent variable
neonatal postnatal hospital stay	#Days from completed delivery till discharge	dependent variable

Continued on the next page.

Variable	**Specification**	**Function**
Treatments		
induced labor	Binary indicator = 1 if induction by membrane sweep, medication, or other procedures (excl. cervical ripening); 0 else	main explanatory variable
non-emergency c-section	Binary indicator if un/planned (excl. emergency) c-section; 0 else	explanatory variable
vaginally operative procedures	Binary indicator = 1 if forceps, spatula, vacuum, episiotomy; 0 else	explanatory variable
Staff capacity instruments		
predicted due date a non-working day	Binary indicator = 1 if predicted due date a Saturday/ Sunday/ public holiday; 0 else (incl. due dates updated by hospital)	instrument
predicted due date not informative	Binary indicator = 1 if hospital discards due date as invalid; 0 else	descriptives
pre-labor membrane rupture	Binary indicator = 1 if condition holds; 0 else	strata
pre-labor membrane rupture at night	Binary indicator = 1 if pre-labor membrane break 8pm-4am; 0 else	instrument
midwife shortage upon admission	0 if no current deliveries; else hospital-minute-wise ratio of #current deliveries w/o midwife/ #all current deliveries	instrument
Obstetrician preferences instruments		
preference induced labor	Obstetrician' s #prior inductions / #all prior deliveries	instrument
preference non-emergency c-section	Obstetrician's #prior non-emergency c-section/ #all prior deliveries	instrument
preference vaginally operative procedures	Obstetrician's #prior vaginally operative procedures/ #all prior deliveries	instrument

Notes: Annual Geburtshilfe datasets provided by the IQTIG institute constitute the main data source supplemented by calendar data to construct the non-working day instrument. A factor variable enters the regression as a set of binary indicators.

Table 3.A.3: Overview Sample Specification

<table>
<tr><td rowspan="15">all births</td><td rowspan="14">non-missing for central variables</td><td rowspan="11">1st births</td><td rowspan="9">MAIN ANALYSIS SAMPLE
zero preconditions</td><td>w pre-labor membrane rupture</td></tr>
<tr><td>at hospitals w/o in-patient midwives</td></tr>
<tr><td>obstetrician unknown pre-admission</td></tr>
<tr><td>mothers admitted after transfer</td></tr>
<tr><td>mothers aged >26</td></tr>
<tr><td>single mothers</td></tr>
<tr><td>mothers w low socioeconomic status</td></tr>
<tr><td>at small hospitals</td></tr>
<tr><td>at low-quality hospitals</td></tr>
<tr><td colspan="2">delivered pre-arrival to hospital</td></tr>
<tr><td colspan="2">others</td></tr>
<tr><td rowspan="2">2nd births</td><td colspan="2">zero preconditions</td></tr>
<tr><td colspan="2">others</td></tr>
<tr><td colspan="3">higher birth orders</td></tr>
<tr><td colspan="4">else</td></tr>
</table>

Table 3.A.4: Characteristics of Births Across Subsamples Dedicated to Heterogeneity Checks

zero-precondition 1st births	**all**	**strata**				
		maternal			hospital	
		age >26	single	low ses	small	low quality
	(1)	(2)	(3)	(4)	(5)	(6)
maternal characteristics						
german (yes/no)	0.81	0.85	0.90	0.80	0.83	0.78
single (yes/no)	0.11	0.10	1.00	0.11	0.12	0.10
low socio-economic status (yes/no)	0.80	0.78	0.77	1.00	0.83	0.76
age	28	30	27	28	28	28
bmi	24	24	24	24	24	24
pre-pregnancy weight (kg)	67	67	68	67	68	67
gestational age (#days)	280	280	280	280	280	280
prenatal care (#doctor visits)	12	12	11	12	12	11
prenatal care starting >12th week (yes/no)	0.077	0.055	0.10	0.081	0.079	0.088
neonatal characteristics						
birth weight (g)	3415	3424	3407	3411	3418	3417
hospital characteristics						
emergency c-section time >20 min (yes/no)	0.015	0.014	0.017	0.015	0.026	0.016
health outcomes						
emergency c-section (yes/no)	0.01	0.01	0.02	0.01	0.01	0.01
perineal tearing (III/IV) (yes/no)	0.02	0.03	0.02	0.02	0.02	0.03
APGAR score (5 min.)	9.7	9.7	9.7	9.7	9.7	9.7
hospital staff capacity outcomes						
labor duration (#hours)	6.8	6.8	6.8	6.7	6.6	6.8
maternal postnatal hospital stay (#days)	3.4	3.4	3.4	3.4	3.5	3.3
neonatal postnatal hospital stay (#days)	3.2	3.2	3.2	3.2	3.4	3.1
treatments						
induced labor (yes/no)	0.28	0.27	0.23	0.28	0.27	0.27
non-emergency c-section (yes/no)	0.21	0.27	0.27	0.26	0.27	0.26
vaginally operative procedues (yes/no)	0.32	0.32	0.32	0.32	0.32	0.32
instruments staff capacity						
predicted due date a non-working day (yes/no)	0.33	0.33	0.34	0.33	0.33	0.34
pre-labor membrane rupture 8pm-4am (yes/no)	0.15	0.16	0.14	0.15	0.14	0.14
midwife shortage upon admission [0,1]	0.58	0.57	0.60	0.58	0.73	0.50
instruments obstetricians' preferences						
preference induced labor [0,1]	0.23	0.23	0.23	0.23	0.23	0.24
preference non-emergency c-section [0,1]	0.34	0.35	0.34	0.35	0.33	0.30
preference vaginally operative procedues [0,1]	0.20	0.20	0.20	0.21	0.21	0.20
N	177,215	119,041	19,986	141,605	91,936	19,914
N obstetricians' preferences	66,916	44,313	2,462	54,194	39,235	10,488

Notes: IQTIG birth records for Germany 2015-2016. Means for the central analysis variables based on zero-precondition 1st births and subsamples stratified by maternal and hospital chracteristics. See Table 3.A.2 and Table 3.A.3 for details on sample and variable construction.

Table 3.A.5: Characteristics of Births Across Subsamples Dedicated to Endogeneity Checks

	zero-precondition 1st births					**delivery pre-arrival only precondition 1st births**
	all	w pre-labor membrane rupture	at hospitals w/o in-patient midwives	unknown to obstetrician pre-admission	admitted after transfer	
	(1)	(2)	(3)	(4)	(5)	(6)
maternal characteristics						
german (yes/no)	0.81	0.81	0.80	0.81	0.78	0.83
single (yes/no)	0.11	0.11	0.14	0.16	0.13	0.12
low socio-economic status (yes/no)	0.80	0.80	0.81	0.79	0.75	0.80
age	28	28	28	28	29	28
bmi	24	24	24	24	24	24
pre-pregnancy weight (kg)	67	67	67	67	67	66
gestational age (#days)	280	278	280	279	280	279
prenatal care (#doctor visits)	12	11	11	11	11	11
prenatal care starting >12th week (yes/no)	0.077	0.072	0.075	0.085	0.087	0.976
neonatal characteristics						
birth weight (g)	3415	3391	3414	3403	3378	3382
hospital characteristics						
emergency c-section time >20 min (yes/no)	0.015	0.013	0.01	0.018	0.018	0.012
health outcomes						
emergency c-section (yes/no)	0.01	0.01	0.02	0.02	0.04	0.02
perineal tearing (III/IV) (yes/no)	0.02	0.02	0.03	0.02	0.02	0.03
APGAR score (5 min.)	9.7	9.7	9.7	9.7	9.1	9.7
hospital staff capacity outcomes						
labor duration (#hours)	6.8	6.7	6.8	6.7	5.7	7.1
maternal postnatal hospital stay (#days)	3.4	3.4	3.4	3.3	3.1	3.2
neonatal postnatal hospital stay (#days)	3.2	3.2	3.2	3.2	2.1	3.1
treatments						
induced labor (yes/no)	0.28	0.35	0.28	0.23	0.33	0.09
non-emergency c-section (yes/no)	0.21	0.23	0.25	0.24	0.34	0.18
vaginally operative procedues (yes/no)	0.32	0.33	0.32	0.33	0.33	0.35
instruments staff capacity						
predicted due date a non-working day (yes/no)	0.33	0.33	0.33	0.33	0.33	0.33
pre-labor membrane rupture 8pm-4am (yes/no)	0.15	0.50	0.16	0.16	0.15	0.26
midwife shortage upon admission [0,1]	0.58	0.57	0.57	0.59	0.65	0.56
instruments obstetricians' preferences						
preference induced labor [0,1]	0.23	0.24	0.25	0.23	0.25	0.23
preference non-emergency c-section [0,1]	0.34	0.34	0.39	0.32	0.35	0.34
preference vaginally operative procedues [0,1]	0.20	0.20	0.21	0.20	0.21	0.20
N	177,215	52,815	64,926	54,198	3,233	4,171
N obstetricians' preferences	66,916	18,885	17,205	17,776	917	1,397

Notes: IQTIG birth records for Germany 2015-2016. Means for the central analysis variables based on zero-precondition 1st births, subsamples, and the sample of mothers without pregnancy or birth risks suffering eclampsia. See Table 3.A.2 and Table 3.A.3 for details on sample and variable construction.

The original OLS model reads

$$Y_m = \underset{1\times t}{\mathbf{t}_m'} \ \underset{t\times 1}{\beta} + \underset{1\times k}{\mathbf{x}_m'} \ \underset{k\times 1}{\delta} + \underset{1\times s}{\lambda} \ \underset{s\times 1}{\mathbf{1}} + v_m \tag{3.3}$$

where

$$\text{treatments } \mathbf{t}_m \equiv \begin{bmatrix} InducedLabor\ (IL_m) \\ CSection\ (CS_m) \\ VaginalOperations\ (VO_m) \\ IL_m * CS_m \\ IL_m * VO_m \\ CS_m * VO_m \\ IL_m * CS_m * VO_m \end{bmatrix}, \text{ covariates } \mathbf{x}_m \equiv \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_{k-1} \end{bmatrix}, \text{ sets of fixed effects } \lambda \equiv \begin{bmatrix} \lambda_1 & \dots & \lambda_s \end{bmatrix}$$

- Y_m: outcome of mother (or her neonate) m
- $IL_m,\ CS_m,\ \ VO_m \in \{0,1\}$: =1 if mother m has an induction (and maybe other interventions), a non-emergency c-section (dto.), or vaginal operations (dto.) respectively; 0 else
- details on pre-determined controls, fixed effects, and cluster-robust standard errors given below Table 3.A.11

Then, the corresponding IV model can be written as

$$\underset{t\times 1}{\mathbf{t}_m} = \underset{t\times z}{\mathbf{\Gamma}} \ \underset{z\times 1}{\mathbf{z}_m} + \underset{t\times k}{\mathbf{\Phi}} \ \underset{k\times 1}{\mathbf{x}_m} + \underset{t\times s}{\mathbf{\Lambda}} \ \underset{s\times 1}{\mathbf{1}} + \underset{t\times 1}{\epsilon_m} \tag{3.4}$$

Notation builds on Equation 3.3. There are $1, ..., t$ treatments, $1, ..., k-1$ covariates, and $1, ..., s$ sets of fixed effects observed for mother m, while $1, ..., z$ instruments are defined as

$$\text{either } \mathbf{z}_m \equiv \begin{bmatrix} DueDateNoWorkday\ (DN_m) \\ MembraneRuptureNight\ (RN_m) \\ MidwifeShortage\ (MS_m) \\ DN_m * RN_m \\ DN_m * MS_m \\ RN_m * MS_m \\ DN_m * RN_m * MS_m \end{bmatrix}, \text{ or } \mathbf{z}_m \equiv \begin{bmatrix} InducedLaborPref\ (ILP_m) \\ CSectionPref\ (CSP_m) \\ VaginalOperPref\ (VOP_m) \\ ILP_m * CSP_m \\ ILP_m * VOP_m \\ CSP_m * VOP_m \\ ILP_m * CSP_m * VOP_m \end{bmatrix}$$

- $DN_m \in \{0,1\}$: =1 if the due date is a weekend day or public holiday, 0 else
- $RN_m \in \{0,1\}$: =1 if mother m has a pre-labor membrane rupture between 8 pm to 4 am, 0 else
- $MS_m \in [0,1] = \begin{cases} 0 \text{ if \#current deliveries at that hospital} = 0 \\ \frac{\text{\#current deliveries at that hospital without a midwife}}{\text{\#current deliveries at that hospital}} \text{ else} \end{cases}$
- $ILP_m,\ CSP_m,\ VOP_m \in [0,1]$: mean prior rate of inductions, non-emergency c-sections, and vaginal operations of obstetrician treating mother m

Table 3.A.6: Multi-Treatment Model Underidentification Tests

	Dependent variable:								
	induced labor	vaginal operations	non-emergency c-section	induction + surgery	non-emergency c-section x vaginal operations	non-emergency c-section x induced labor	vaginal operations x induced labor	non-emergency c-section x vaginal operations x induced labor	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
obstetricians' intervention preferences									66,916
	0.0000	0.0000	0.0000	0.0000					
	0.2800	0.6400	0.9200		0.7500	0.4100	0.5000	1.0000	
hospital staffing constraints									177,215
	0.8800	0.8000	0.7700	0.8000					
	0.9900	0.9500	0.7400		0.9500	1.0000	0.6900	1.0000	
Mean (dep. var.)	0.2770	0.3161	0.2558		0.0003	0.0896	0.0820	0.0001	

Notes: *p<0.1; **p<0.05; ***p<0.01. P-values reported for Sanderson and Windmeijer (2016)'s cluster-robust underidentification test (H_0 : *There is underidentification.*) for multiple endogenous treatments derived from running the Sargan-Hansen J-Test for overidentification (Cameron and Miller, 2015) in auxiliary regressions. The 1st set of instruments is derived from obstretricians' preferences to perform induction, c-section, or vaginal oerations. The 2nd set, based on hospital staffing constraints, involves *Midwife shortages upon arrival*, *Prelabor membrane rupture during night shift*, and *Due date a non-working day.* To identify a four-treatment model (Equation 3.1), the instruments' triple interaction is added to each set; for seven-treatment models (Equation 3.3), all instruments' interactions are added. The auxiliary regressions incl. core controls and are based on zero-precondition first-births (the 1st instrument set restricts further to non-missing obstetrician ids , see Table 3.A.2 and Table 3.A.3 for details on sample and variable construction). Robust standard errors clustered by 3-digit zip codes of maternal residence. Means are only available for the full sample of zero-precondition first births.

Table 3.A.7: First-Stage Effects Based on Hospital Staff Capacity Constraints

	Dependent variable:						
	induced labor	vaginally operative procedures	non-emergency c-section	non-emergency c-section x vaginally operative procedures	non-emergency c-section x induced labor	vaginally operative procedures x induced labor	non-emergency c-section x vaginally operative procedures x induced labor
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Midwife shortage upon admission	*insign.*	*sign.*	*sign.*	*insign.*	*sign.*	*insign.*	*insign.*
Due date non-working day	*insign.*	*insign.*	*insign.*	*insign.*	*insign.*	*insign.*	*insign.*
Pre-labor membrane rupture at night	*sign.*	*sign.*	*sign.*	*insign.*	*insign.*	*sign.*	*sign.*
Midwife shortage upon admission x due date non-working day	*sign.*	*insign.*	*insign.*	*insign.*	*insign.*	*insign.*	*insign.*
Midwife shortage upon admission x pre-labor membrane rupture at night	*sign.*	*insign.*	*insign.*	*insign.*	*insign.*	*insign.*	*insign.*
Due date non-working day x pre-labor membrane rupture at night	*insign.*	*insign.*	*insign.*	*insign.*	*insign.*	*sign.*	*insign.*
Midwife shortage upon admission x due date non-working day x pre-labor membrane rupture at night	*insign.*	*insign.*	*insign.*	*insign.*	*insign.*	*sign.*	*insign.*
Mean (dependent variable)	0.2770	0.3161	0.2558	0.0003	0.0896	0.0820	0.0001
Underidentification (p-value)	0.9900	0.9500	0.7400	0.9500	1.0000	0.6900	1.0000

Notes: *p<0.1; **p<0.05; ***p<0.01. IQTIG birth records for Germany 2015-2016. The main sample are zero-precondition first-births (N=177,215). Sample and variable creation detailed in Table 3.A.2 and Table 3.A.3. *For some remote execution issue, coeffcients on interacted dependent variables are not released, only an indicator for at least 10%-level significance "sign." or less "insign.".* Underlying regressions follow model Equation 3.3 and use robust standard errors clustered by 3-digit zip codes of maternal residence. Underidentification is tested (see Table 3.A.6 for details). Means are available for the main sample.

Table 3.A.8: First-Stage Effects Based on Obstetricians' Preferences for Interventions

	Dependent variable:						
	induced labor	vaginally operative procedures	non-emergency c-section	non-emergency c-section x vaginally operative procedures	non-emergency c-section x induced labor	vaginally operative procedures x induced labor	non-emergency c-section x vaginally operative procedures x induced labor
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Induced labor preference	*sign.*	*sign.*	*insign.*	*sign.*	*insign.*	*sign.*	*sign.*
Non-emergency c-section preference	*sign.*	*insign.*	*sign.*	*insign.*	*sign.*	*sign.*	*sign.*
Vaginally operative procedures' preference	*sign.*	*sign.*	*sign.*	*insign.*	*sign.*	*sign.*	*sign.*
Induced labor preference x Non-emergency c-section preference	*insign.*	*insign.*	*insign.*	*insign.*	*sign.*	*insign.*	*sign.*
Induced labor preference x Vaginally operative procedures' preference	*sign.*	*insign.*	*insign.*	*insign.*	*sign.*	*insign.*	*sign.*
Non-emergency c-section preference x Vaginally operative procedures' preference	*sign.*	*sign.*	*insign.*	*insign.*	*sign.*	*insign.*	*sign.*
Induced labor preference x Non-emergency c-section preference x Vaginally operative procedures' preference	*sign.*	*insign.*	*insign.*	*sign.*	*insign.*	*sign.*	*insign.*
Mean (dependent variable)	0.2770	0.3161	0.2558	0.0003	0.0896	0.0820	0.0001
Underidentification (p-value)	0.2800	0.6400	0.9200	0.7500	0.4100	0.5000	1.0000

Notes: *p<0.1; **p<0.05; ***p<0.01. IQTIG birth records for Germany 2015-2016. The main sample are zero-precondition first-births (N=177,215). Instrumenting by intervention preferences creates a subsample of births with non-missing obstetrician id (N=66,916). Sample and variable creation detailed in Table 3.A.2 and Table 3.A.3. *For some remote execution issue, coefficents on interacted dependent variables are not released, only an indicator for at least 10%-level significance "sign." or less "insign.".* Underlying regressions follow model Equation 3.3 and use robust standard errors clustered by 3-digit zip codes of maternal residence. Underidentification is tested (see Table 3.A.6 for details). Means are available for the main sample.

Table 3.A.9: Reduced Form Effects Based on Hospital Staff Capacity Constraints

	Dependent variable:				
	maternal health		staff capacity		
	emergency c-section	perineal tearing (III/IV)	labor duration (#hours)	hospital stay	
				mother	neonate
	(1)	(2)	(3)	(4)	(5)
Midwife shortage upon admission	0,0024***	-0,0027***	-1,0266***	0,2289***	0,2439***
	(0,0009)	(0,0010)	(0,0597)	(0,0159)	(0,0184)
Due date non-working day	-0.0009	0.0006	0.0622	0.0157	0.0406**
	(0.0009)	(0.0013)	(0.0489)	(0.0130)	(0.0180)
Pre-labor membrane rupture at night	-0.0007	-0.0011	-0.1124	-0.0392**	-0.0317
	(0.0013)	(0.0020)	(0.0682)	(0.0179)	(0.0204)
Midwife shortage upon admission x Due date non-working day	-0.0002	-0.0003	-0.0556	-0.0271	-0.0410
	(0.0012)	(0.0017)	(0.0629)	(0.0174)	(0.0250)
Midwife shortage upon admission x Pre-labor membrane rupture at night	-0.0030*	0.0016	0.1553*	-0.0034	-0.0240
	(0.0017)	(0.0024)	(0.0878)	(0.0249)	(0.0266)
Due date non-working day x Pre-labor membrane rupture at night	0.0007	-0.0034	0.0108	-0.0116	-0.0161
	(0.0023)	(0.0033)	(0.1128)	(0.0282)	(0.0348)
Midwife shortage upon admission x Due date non-working day x Pre-labor membrane rupture at night	0.0046	0.0022	-0.1053	0.0465	0.0761
	0.0032	(0.0042)	(0.1547)	(0.0406)	(0.0567)
Mean (dependent variable)	0.01	0.02	6.80	3.40	3.20
Adjusted R^2	0.0011	0.0017	0.0147	0.0162	0.0079

Notes: *p<0.1; **p<0.05; ***p<0.01. IQTIG birth records for Germany 2015-2016. The main sample are zero-precondition first-births (N=177,215). Sample and variable creation detailed in Table 3.A.2 and Table 3.A.3. Reported coefficients stem from reduced forms following model Equation 3.3 (detailed below Table 3.A.11). Robust standard errors clustered by 3-digit zip codes of maternal residence. Adjusted R^2 reported for regressions including core controls.

Table 3.A.10: Reduced Form Effects Based on Obstetricians' Preferences for Interventions

	Dependent variable:				
	maternal health		staff capacity		
	emergency c-section	perineal tearing (III/IV)	labor duration (#hours)	hospital stay	
				mother	neonate
	(1)	(2)	(3)	(4)	(5)
Induced labor preference	-0.0004 (0.0039)	0.0040 (0.0095)	-1.0944*** (0.3743)	-0.2074* (0.1065)	-0.2102** (0.1058)
Non-emergency c-section preference	0.0401*** (0.0057)	0.0070 (0.0048)	-5.0095*** (0.2763)	0.9834*** (0.0562)	0.7371*** (0.0801)
Vaginally operative procedures' preference	0.0128** (0.0056)	0.0026 (0.0084)	-1.7420*** (0.3457)	0.6928*** (0.0989)	0.6602*** (0.1156)
Induced labor preference x Non-emergency c-section preference	0.0432** (0.0206)	0.0004 (0.0161)	1.6491** (0.7761)	0.1917 (0.1742)	0.1332 (0.2101)
Induced labor preference x Vaginally operative procedures' preference	0.0017 (0.0123)	-0.0111 (0.0226)	2.1683** (0.8737)	-0.0396 (0.2014)	0.0654 (0.2644)
Non-emergency c-section preference x Vaginally operative procedures' preference	-0.0638** (0.0287)	0.0507 (0.0314)	6.3833*** (1.4891)	0.8410** (0.3739)	1.1898** (0.4624)
Induced labor preference x Non-emergency c-section preference x Vaginally operative procedures' preference	0.1527 (0.0982)	0.0993 (0.0991)	-8.7465** (4.4016)	-1.6280 (1.1039)	-2.7907** (1.2280)
Mean (dependent variable)	0.01	0.02	6.80	3.40	3.20
Adjusted R^2	0.0092	0.0025	0.0402	0.0433	0.0157

Notes: *p<0.1; **p<0.05; ***p<0.01. IQTIG birth records for Germany 2015-2016. The main sample are zero-precondition first-births. Instrumenting by intervention preferences creates a subsample of births with non-missing obstetrician id (N=66,916). Sample and variable creation detailed in Table 3.A.2 and Table 3.A.3. Reported coefficients stem from reduced forms following model Equation 3.3 (detailed below Table 3.A.11). Robust standard errors clustered by 3-digit zip codes of maternal residence. Adjusted R^2 reported for regressions including core controls. Means are available for the main sample.

Table 3.A.11: Maternal Health Effects of Non-Medically Indicated Induced Labor

	OLS		IV			
Instruments based on			staff capacity		obstetricians' preferences	
Dependent variable	emergency c-section	perineal tearing (III/IV)	emergency c-section	perineal tearing (III/IV)	emergency c-section	perineal tearing (III/IV)
	(1)	(2)	(3)	(4)	(5)	(6)
No controls	0.0205***	0.0029**	0.2594	-0.1425	-0.2074	0.0867
	(0.0017)	(0.0013)	(6.5993)	(0.3221)	(0.3652)	(0.5341)
Core controls	0.0201***	0.0032**	-0.7842	-0.1189	-0.1710	0.1017
	(0.0017)	(0.0013)	(53.4963)	(1.3234)	(0.3484)	(0.5327)
Add month, weekday, hour FE	0.0197***	0.0032**	0.3764	-0.1283	-0.2067	0.1176
	(0.0018)	(0.0013)	(4.9598)	(0.2761)	(0.2255)	(0.3652)
Add hospital controls	0.0199***	0.0025*	0.4959	-0.1255	-0.1829	0.0523
	(0.0018)	(0.0013)	(8.6641)	(0.4045)	(0.2214)	(0.3540)
Add hospital FE	0.0195***	0.0025*	0.2141	-0.1555	-0.4820	0.4643
	(0.0018)	(0.0013)	(0.2688)	(0.2232)	(0.9310)	(0.6601)
Core controls & labor	0.0224***	0.0033**	0.1021	-0.1333	-0.3167	0.0420
	(0.0017)	(0.0013)	(0.1990)	(0.1303)	(0.3714)	(0.6500)
Miscelleanous controls	0.0200***	0.0030**	0.1776	-0.1466	-0.1081	-0.0255
	(0.0017)	(0.0013)	(0.8141)	(0.1343)	(0.3033)	(0.4903)
Main effects only (core controls)	0.0081***	0.0029***	0.0996	-0.0563	0.0730***	0.0223
	(0.0007)	(0.0009)	(0.1968)	(0.1920)	(0.0196)	(0.0207)
Mean (dependent variable)	0.01	0.02	0.01	0.02		
Adjusted R^2	0.0227	0.0157				
N	177,215	177,215	177,215	177,215	66,916	66,916

Notes: *p<0.1; **p<0.05; ***p<0.01. Birth records from all German hospitals covering 2015-2016, provided by the IQTIG institute. Linear probability models based on the main analysis sample of zero-precondition first births (neither pregnancy nor birth risks known antepartum), for which all central regression inputs are non-missing. Sample and variable creation detailed in Table 3.A.2 and Table 3.A.3. Reported coefficients for induced labor stem from separate regressions following model Equation 3.3. Intervention treatments are represented by binary indicators for induced labor, non-emergency c-sections, and vaginally operative procedures. Staff capacity-based instruments are binary indicators for a mother's due date on a non-working day, a pre-labor rupture of membranes between 8 pm and 4 am, and a minute-wise measure $\in [0,1]$ of midwife shortages upon maternal admission. The instruments based on obstetricians' preferences are computed for each of the three main interventions as the mean intervention rate across an obstetrician's past deliveries. Instrumenting by intervention preferences creates a subsample of births with non-missing obstetrician id. Treatments and instruments enter as main effects and interactions. Core controls include the year of delivery, a mother's age, her region of origin (7 categories), her socio-economic status (6 categories), and her single status (yes/no), where categorical variables enter as sets of binary indicators. Moreover, continuous measures are created for maternal height (as cubic), maternal weight at the beginning of the pregnancy (as cubic), and maternal BMI. Binary hospital (stay) controls indicate whether 1) the mother brings her own midwife, 2) she has been introduced to her obstetrician during pregnancy, and 3) documentation of her delivery seem to be done in a haste. Miscellaneous controls are binary indicators for maternal alcohol consumption, psychological or social problems, minor diseases or pregnancy risks , as well the count of doctor visits. Finally, there is a dummy for maternal employment status. Main effects only (...) refers to the core specification w/o interactions of treatments or instruments. Robust standard errors clustered by 3-digit zip code of maternal residence. Adjusted R^2 reported for regressions including core controls. Means are available for the main sample.

Table 3.A.12: Sample-specific Health Effects of Non-Medically Indicated Induced Labor

Instruments based on	staff capacity		obstetricians' preferences			
Dependent variable	emergency c-section	perineal tearing (III/IV)	emergency c-section	perineal tearing (III/IV)	$N_{(1)-(2)}$	$N_{(3)-(4)}$
	(1)	(2)	(3)	(4)		
zero-precondition 1st births	-0.7842	-0.1189	-0.1710	0.1017	177,215	66,916
	(53.4963)	(1.3234)	(0.3484)	(0.5327)		
w pre-labor membrane rupture	0.0613	0.1401	0.3216	0.1087	52,815	18,885
	(7.6489)	(8.0396)	(0.3645)	(0.1642)		
at hospitals w/o in-patient midwives	-0.9998	0.3518	0.0147	0.0610	64,926	17,205
	(11.0254)	(2.3171)	(0.1188)	(0.1712)		
unknown to obstetrician pre-admission	-0.1817	-0.2840	0.0485	-0.1975	54,198	17,776
	(1.2580)	(1.2355)	(0.6325)	(0.4618)		
admitted after transfer	0.3296	-0.0541	0.6262	0.2765	3,233	917
	(0.5707)	(0.3147)	(1.2853)	(1.4799)		
to mothers aged >26	0.3738	-0.1431	0.8961	-1.9978	119,041	44,313
	(1.2845)	(0.7605)	(2.5987)	(5.5746)		
to single mothers	-0.0081	-0.3490	0.1557	0.1207	19,986	2,462
	(0.2146)	(0.1988)	(0.2375)	(0.2403)		
to mothers w low socio-economic status	0.0383	-0.1341	0.0140	-0.4802	141,605	54,194
	(0.8211)	(0.3339)	(0.3694)	(0.9299)		
at small hospitals	0.0771	-0.0767	-0.9108	0.2442	91.936	39.235
	(0.3858)	(0.2856)	(1.2815)	(0.6612)		
at low quality hospitals	-0.2210	-0.0073	-0.1292	-0.0496	19,914	10,488
	(0.3938)	(0.3432)	(0.1752)	(0.1527)		
delivery pre-arrival	2.0004	-14.9676	0.0752	-0.3553	4,171	1,395
	(83.8723)	(572.1791)	(0.8968)	(0.4475)		
zero-precondition 2nd births	-0.0542	0.0044	0.0391	-0.0310	81,896	27,558
	(0.1918)	(0.1239)	(0.0414)	(0.0337)		

Notes: *p<0.1; **p<0.05; ***p<0.01. IQTIG birth records for Germany 2015-2016. The main sample are zero-precondition first-births. Instrumenting by intervention preferences creates a subsample of births with non-missing obstetrician id. Sample and variable creation detailed in Table 3.A.2 and Table 3.A.3. Reported coefficients for induced labor stem from separate regressions following model Equation 3.3 (detailed below Table 3.A.11) but run for alternative samples (see Table 3.A.3). Robust standard errors clustered by 3-digit zip codes of maternal residence.

Table 3.A.13: Hospital Staff Capacity Effects of Non-Medically Indicated Induced Labor

	OLS			IV					
Instruments based on				staff capacity			obstetricians' preferences		
Dependent variable	labor duration (#hours)	postnatal hospital stay (#days)		labor duration (#hours)	postnatal hospital stay (#days)		labor duration (#hours)	postnatal hospital stay (#days)	
		mother	neonate		mother	neonate		mother	neonate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
No controls	-1.0353***	0.1415***	0.0989***	-13.5044	1.1960	2.8748	-29.7075	3.7541	2.7828
	(0.0498)	(0.0110)	(0.0135)	(373.9101)	(14.0921)	(14.3149)	(47.3025)	(5.6907)	(5.9303)
Core controls	-1.0117***	0.1432***	0.1068***	44.4728	-0.8431	0.6261	-31.8266	3.8466	2.9638
	(0.0500)	(0.0109)	(0.0134)	(2986.7851)	(111.1334)	(128.8600)	(51.4751)	(5.9850)	(6.2079)
Add month, weekday, hour FE	-1.0231***	0.0883***	0.0509***	-29.7799	2.1956	4.1533	-20.5187	2.1190	1.6719
	(0.0511)	(0.0107)	(0.0141)	(464.6725)	(26.0612)	(21.4955)	(32.0754)	(3.3148)	(3.8279)
Add hospital controls	-1.0371***	0.0911***	0.0554***	-41.0236	2.6662	4.4217	-28.3986	2.0995	0.7391
	(0.0513)	(0.0107)	(0.0141)	(809.5115)	(40.8639)	(30.3282)	(33.1837)	(3.6160)	(3.8385)
Add hospital FE	-0.9030***	0.1168***	0.0920***	-8.9386	0.1634	1.8641	-12.6147	4.2166	-0.3783
	(0.0474)	(0.0105)	(0.0135)	(7.7969)	(2.8216)	(5.9000)	(33.3624)	(5.9394)	(5.4054)
Core controls & labor	-1.2397***	0.1451***	0.1074***	-6.6251	1.0142	2.8021	-26.5807	3.5885	2.9673
	(0.0507)	(0.0109)	(0.0134)	(8.7821)	(1.5816)	(2.7709)	(43.2773)	(5.2765)	(5.5831)
Miscelleanous controls	-1.0476***	0.1471***	0.1169***	-9.8327	1.0803	2.8048	-39.1492	3.6921	2.0941
	(0.0514)	(0.0108)	(0.0135)	(53.4313)	(2.3680)	(3.7456)	(50.1474)	(5.8272)	(5.4258)
Main effects only (core controls)	-0.7063***	0.0897***	0.0452***	-7.4338	-0.5037	-5.0500	-1.6727	-1.1620***	-1.3736***
	(0.0401)	(0.0080)	(0.0115)	(11.2756)	(3.6683)	(9.5341)	(1.0939)	(0.2981)	(0.3351)
Mean (dependent variable)	6.80	3.40	3.20	6.80	3.40	3.20			
Adjusted R^2	0.1459	0.1236	0.0568						
N	177,215	177,215	177,215	177,215	177,215	177,215	177,215	66,916	66,916

Notes: *p<0.1; **p<0.05; ***p<0.01. IQTIG birth records for Germany 2015-2016. The main sample are zero-precondition first-births. Instrumenting by intervention preferences creates a subsample of births with non-missing obstetrician id. Sample and variable creation detailed in Table 3.A.2 and Table 3.A.3. Reported coefficients for induced labor stem from separate regressions following model Equation 3.3 (detailed below Table 3.A.11). Robust standard errors clustered by 3-digit zip codes of maternal residence. Adjusted R^2 reported for regressions including core controls. Means are available for the main sample.

Table 3.A.14: Relative Effects of Induction vs. Surgical Intervention

	Dependent variable: hospital stay (# days)					
	OLS		IV			
			staff capacity		obstetricians' preferences	
	mother (1)	neonate (2)	mother (3)	neonate (4)	mother (5)	neonate (6)
Induced labor	0,1432*** (0.0109)	0.1068*** (0.0134)	-0.8431 (111.1334)	0.6261 (128.8600)	3.8466 (5.9850)	2.9638 (6.2079)
Non-emergency c-section	1.3202*** (0.0175)	1.1537*** (0.0198)	-9.8964 (752.6946)	-11.4455 (868.8066)	2.3109** (1.1644)	2.0986* (1.1971)
Vaginal operations	0.1828*** (0.0114)	0.1493*** (0.0139)	-29.3680 (1942.0808)	-31.5018 (2240.6533)	4.4629 (3.8845)	3.9694 (4.1319)
Induced labor x Non-emergency c-section	-0.1191*** (0.0192)	-0.1239*** (0.0292)	69.0230 (3863.9451)	76.0869 (4459.4136)	-3.3350 (4.8581)	-3.5243 (4.9821)
Induced labor x Vaginal operations	-0.0681*** (0.0174)	-0.0943*** (0.0220)	9.3691 (759.7923)	6.6078 (875.8903)	-12.8115 (14.8890)	-10.1293 (15.6703)
Non-emergency c-section x Vaginal operations	-0.2544 (0.1984)	-0.2741 (0.2528)	-107.0354 (2575.7215)	-240.5277 (2882.4232)	-130.1122 (631.5509)	-281.2032 (670.7902)
Induced labor x Non-emerg. c-section x Vag. oper.	0.0907 (0.3656)	0.2971 (0.4860)	2686.9309 (154104.5424)	3597.2043 (177408.8657)		
Mean (dependent variable)	3.40	3.20	3.40	3.20		
Adjusted R^2	0.1236	0.0568				
N	177,2150	177,2150	177,2150	177,2150	66,916	66,916

Notes: *p<0.1; **p<0.05; ***p<0.01. IQTIG birth records for Germany 2015-2016. The main sample are zero-precondition first-births. Instrumenting by intervention preferences creates a subsample of births with non-missing obstetrician id. Sample and variable creation detailed in Table 3.A.2 and Table 3.A.3. Reported coefficients for intervention follow model Equation 3.3 (detailed below Table 3.A.11). Robust standard errors clustered by 3-digit zip codes of maternal residence. Adjusted R^2 reported for regressions including core controls. Means are available for the main sample.

Table 3.A.15: Sample-specific Staff Capacity Effects of Non-Medically Indicated Induced Labor

Instruments based on	staff capacity			obstetricians' preferences				
Dependent variable	labor duration (#hours)	postnatal hospital stay (#days)		labor duration (#hours)	postnatal hospital stay (#days)			
		mother	neonate		mother	neonate	$N_{(1)-(3)}$	$N_{(4)-(6)}$
	(1)	(2)	(3)	(4)	(5)	(6)		
zero-precondition 1st births	44.4728	-0.8431	0.6261	-31.8266	3.8466	2.9638	177,215	66,916
	(2986.7851)	(111.1334)	(128.8600)	(51.4751)	(5.9850)	(6.2079)		
with pre-labor membrane rupture	13.2418	31.4363	53.7219	-7.1686	-1.1150	-0.1051	52,815	18,885
	(407.1023)	(604.7513)	(1065.8878)	(8.9018)	(2.5213)	(3.2961)		
at hospitals w/o in-patient midwives	73.7779	-22.1140	-12.9251	4.5229	0.2519	0.9101	64,926	17,205
	(626.3206)	(174.6608)	(77.2291)	(6.6268)	(2.1979)	(2.2216)		
unknown to obstetrician pre-admission	-6.3800	-8.8533	-5.5982	33.4803	-3.2455	-2.2927	54,198	17,776
	(25.7600)	(30.5830)	(27.8711)	(27.9019)	(5.2948)	(5.9328)		
admitted after transfer	6.2873	1.6719	-1.3041	22.8765	7.2392	11.8903	3,233	917
	(21.1890)	(5.1644)	(2.5281)	(42.6121)	(13.5778)	(16.6935)		
to mothers aged >26	-15.7227	0.4123	2.9843	21.5082	-12.4561	-12.0063	119,041	44,313
	(80.5652)	(5.2204)	(11.1776)	(167.0019)	(30.7977)	(29.4835)		
to single mothers	5.9564	-0.9699	-0.6513	4.9491	3.8176	3.2246	19,986	2,462
	(17.8985)	(4.0973)	(3.3721)	(12.6889)	(3.5791)	(4.0644)		
to mothers w low socio-economic status	2.1151	0.5567	1.6434	-70.7439	4.9493	0.2901	141,605	54,194
	(41.6917)	(2.9570)	(3.5705)	(100.3852)	(8.9630)	(5.5070)		
at small hospitals	2.5659	-2.8901	-0.2278	-27.0442	9.1423	2.8971	91,936	39,235
	(36.5997)	(4.2561)	(12.4331)	(39.6827)	(18.8312)	(13.7552)		
at low-quality hospitals	-32.0840	6.1798	3.4350	14.2136	1.7202	4.1716	19,914	10,488
	(27.4157)	(8.8278)	(9.6531)	(14.8235)	(1.7062)	(3.6793)		
delivery pre-arrival	-590.1633	-121.3088	-88.6764	24.5799	-9.3640	-6.5620	4,171	1,395
	(26621.1444)	(4140.5791)	(2906.5923)	(30.8458)	(7.9748)	(7.1380)		
zero-precondition 2nd births	-2.1172	-3.6416	-0.2685	-5.1756	0.5379	-0.1112	81,896	27,558
	(12.0511)	(5.1512)	(4.6309)	(1.7290)	(0.7955)	(0.7917)		

Notes: *p<0.1; **p<0.05; ***p<0.01. IQTIG birth records for Germany 2015-2016. The main sample are zero-precondition first-births. Instrumenting by intervention preferences creates a subsample of births with non-missing obstetrician id. Sample and variable creation detailed in Table 3.A.2 and Table 3.A.3. Reported coefficients for induced labor stem from separate regressions following model Equation 3.3 (detailed below Table 3.A.11) but run for alternative samples (see Table 3.A.3). Robust standard errors clustered by 3-digit zip codes of maternal residence.

Table 3.A.16: Relative Effects of Induction vs. Surgical Intervention

	Dependent variable: hospital stay (# days)					
	OLS		IV			
			staff capacity		obstetricians' preferences	
	mother (1)	neonate (2)	mother (3)	neonate (4)	mother (5)	neonate (6)
Induced labor	0,1432*** (0.0109)	0.1068*** (0.0134)	-0.8431 (111.1334)	0.6261 (128.8600)	3.8466 (5.9850)	2.9638 (6.2079)
Non-emergency c-section	1.3202*** (0.0175)	1.1537*** (0.0198)	-9.8964 (752.6946)	-11.4455 (868.8066)	2.3109** (1.1644)	2.0986* (1.1971)
Vaginal operations	0.1828*** (0.0114)	0.1493*** (0.0139)	-29.3680 (1942.0808)	-31.5018 (2240.6533)	4.4629 (3.8845)	3.9694 (4.1319)
Induced labor x Non-emergency c-section	-0.1191*** (0.0192)	-0.1239*** (0.0292)	69.0230 (3863.9451)	76.0869 (4459.4136)	-3.3350 (4.8581)	-3.5243 (4.9821)
Induced labor x Vaginal operations	-0.0681*** (0.0174)	-0.0943*** (0.0220)	9.3691 (759.7923)	6.6078 (875.8903)	-12.8115 (14.8890)	-10.1293 (15.6703)
Non-emergency c-section x Vaginal operations	-0.2544 (0.1984)	-0.2741 (0.2528)	-107.0354 (2575.7215)	-240.5277 (2882.4232)	-130.1122 (631.5509)	-281.2032 (670.7902)
Induced labor x Non-emerg. c-section x Vag. oper.	0.0907 (0.3656)	0.2971 (0.4860)	2686.9309 (154104.5424)	3597.2043 (177408.8657)		
Mean (dependent variable)	3.40	3.20	3.40	3.20		
Adjusted R^2	0.1236	0.0568				
N	177,2150	177,2150	177,2150	177,2150	66,916	66,916

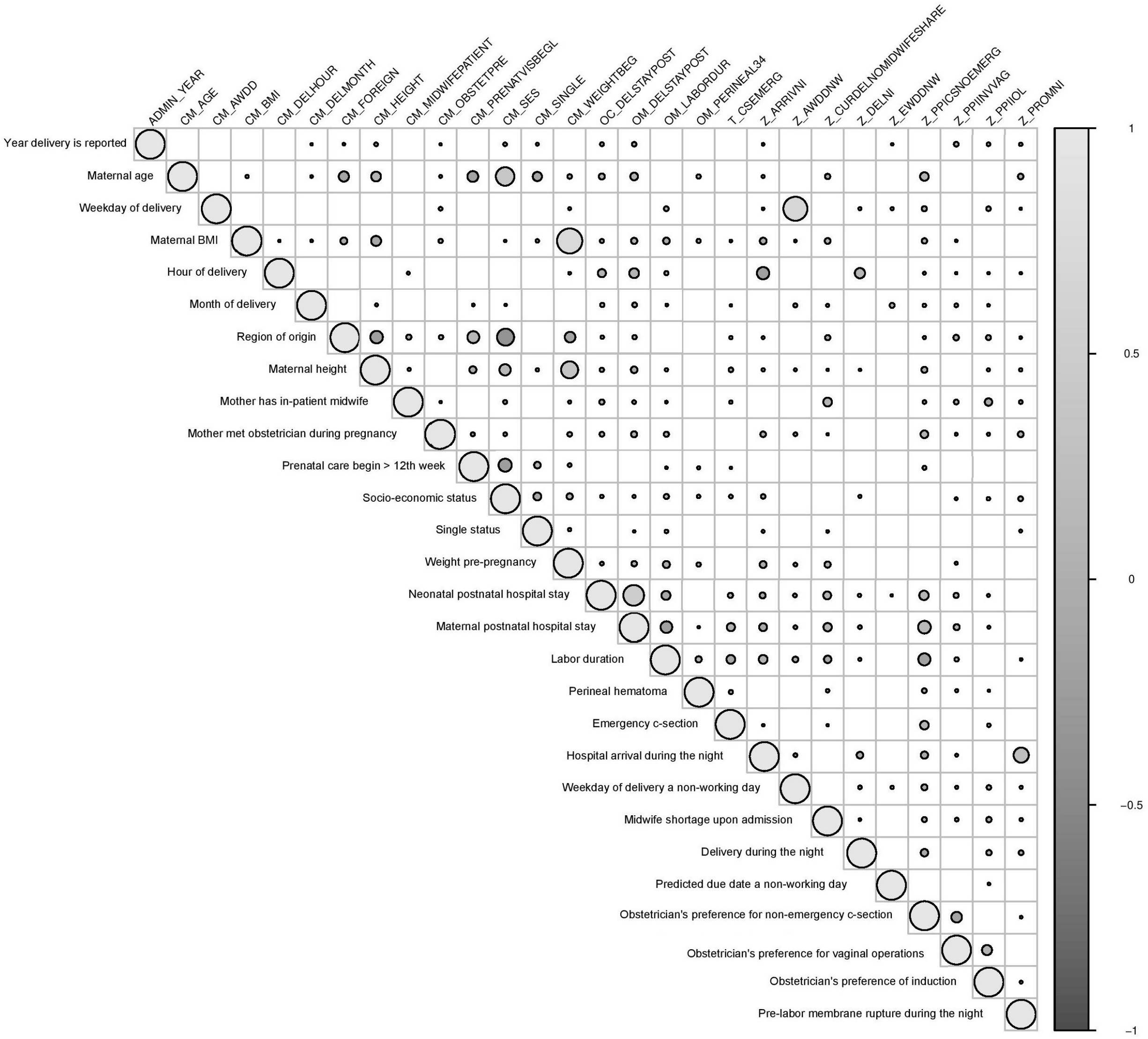
Notes: *p<0.1; **p<0.05; ***p<0.01. IQTIG birth records for Germany 2015-2016. The main sample are zero-precondition first-births. Instrumenting by intervention preferences creates a subsample of births with non-missing obstetrician id. Sample and variable creation detailed in Table 3.A.2 and Table 3.A.3. Reported coefficients for intervention follow model Equation 3.3 (detailed below Table 3.A.11). Robust standard errors clustered by 3-digit zip codes of maternal residence. Adjusted R^2 reported for regressions including core controls. Means are available for the main sample.

Table 3.A.17: Placebo Effects of Non-Medically Indicated Induced Labor

	OLS	IV	
Instruments based on		staff capacity	obstetricians' preferences
Dependent variable	1st prenatal care >12th week		
	(1)	(2)	(3)
No controls	0.0071*** (0.0024)	-0.0459 (2.3953)	0.9548 (1.0697)
Core controls	0.0006 (0.0023)	0.6016 (31.7039)	0.9232 (1.1911)
Add month. weekday. hour FE	0.0016 (0.0024)	-0.2098 (4.3706)	0.5529 (0.6172)
Add hospital controls	0.0024 (0.0024)	-0.3335 (8.0282)	0.5504 (0.6471)
Add hospital FE	0.0035 (0.0024)	0.0041 (0.3108)	0.6069 (1.0568)
Core controls & labor	0.0006 (0.0023)	0.0358 (0.3282)	0.7703 (1.0084)
Miscellaneous controls	0.0081*** (0.0023)	-0.0023 (0.7882)	0.5192 (0.7658)
Main effects only (core controls)	0.0025* (0.0014)	0.0520 (0.3512)	-0.0397 (0.0394)
Mean (dependent variable)	0.077	0.077	
Adjusted R^2	0.0636		
N	177,215	177,215	66,916

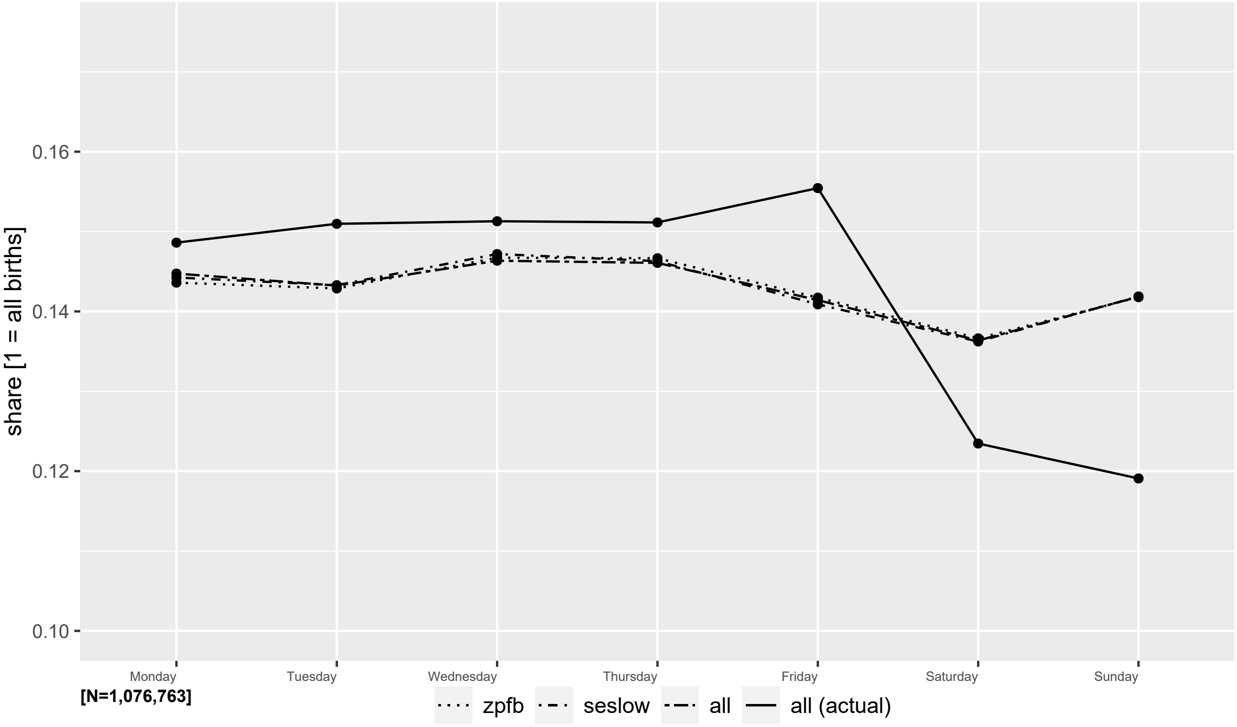
Notes: *p<0.1; **p<0.05; ***p<0.01. IQTIG birth records for Germany 2015-2016. The main sample are zero-precondition first-births. Instrumenting by intervention preferences creates a subsample of births with non-missing obstetrician id. Sample and variable creation detailed in Table 3.A.2 and Table 3.A.3. Reported coefficients for induced labor stem from separate regressions following model Equation 3.3 (detailed below Table 3.A.11). Robust standard errors clustered by 3-digit zip codes of maternal residence. Adjusted R^2 reported for regressions including core controls. Means are available for the main sample.

Figure 3.A.1: Heatmap of Unconditional Correlations of Central Variables



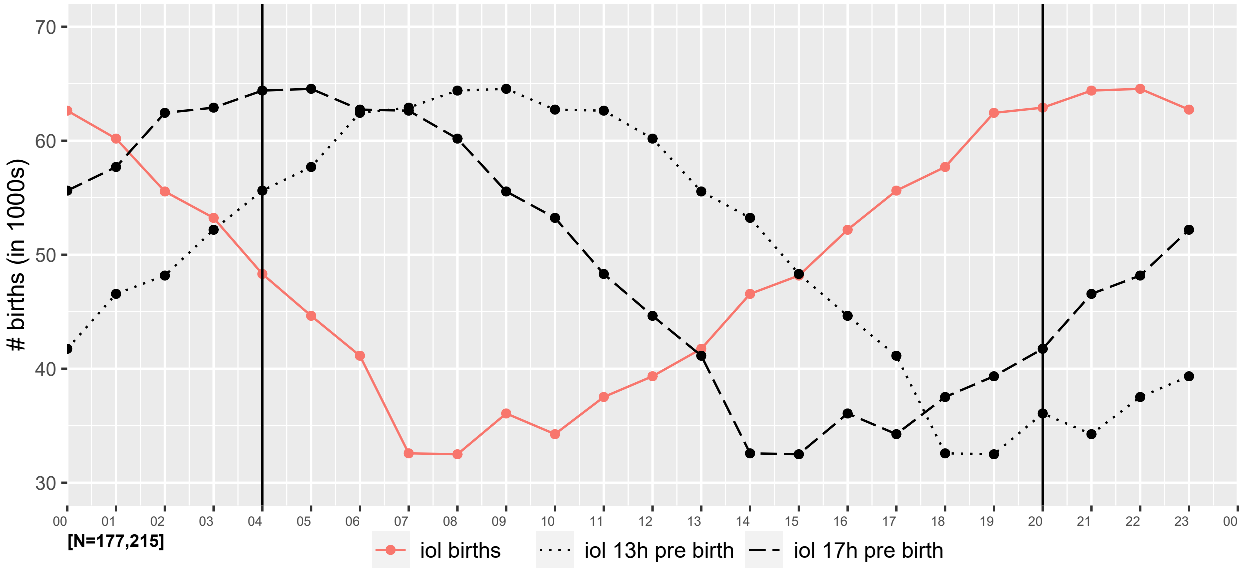
Source: IQTIG German hospital birth records for 2015-2016 restricted to a subsample of zero-precondition first births with information on obstetrician ids (N=66,916 out of 177,215). Sample and variable construction is detailed in the notes to Table 3.A.2 and Table 3.A.3. Variables not derived from obstetrician ids correlate similarly in the main sample. Correlation *values* are indicated by the color ramp, *significance* by the size of circles (no circle if insignificant at the 10% level). Variable names, not labels are shown on the horizontal line. Own calculations.

Figure 3.A.2: Socioeconomic Status & Due Date Distribution Across Weekdays



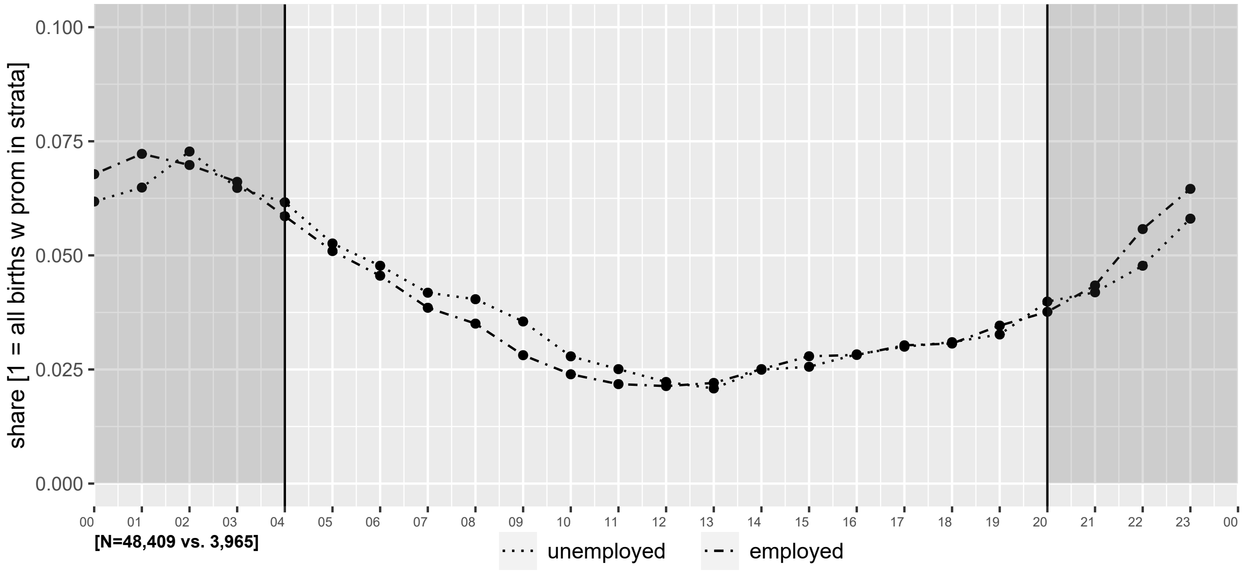
Source: IQTIG German hospital birth records for 2015-2016. *all* refers to the sample of all 1st and 2nd births. *zpfb* refers to the main analysis sample of zero-precondition first births (detailed in Table 3.A.2 and Table 3.A.3), *seslow* restricts this sample to mothers with lower socioeconomic status. Benchmarking fluctuations in predicted due dates, *actual* plots actual weekdays of delivery. Own calculations.

Figure 3.A.3: Intervention & Delivery Timing of Induced Births Across Daily Hours



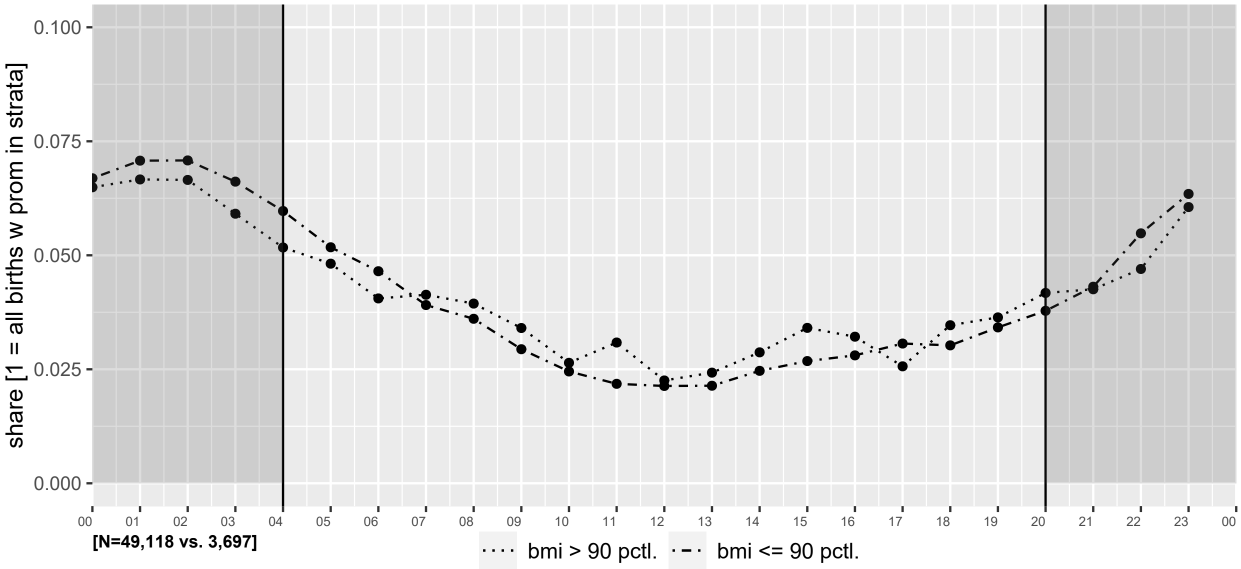
Source: IQTIG German hospital birth records for 2015-2016 restricted to the main analysis sample of zero-precondition first births (detailed in Table 3.A.2 and Table 3.A.3). Unobserved induction timing is proxied lagging birth timing by, e.g., 17 hours. Own calculations.

Figure 3.A.4: Employment Status & Births after Pre-Labor Membrane Ruptures Across Daily Hours



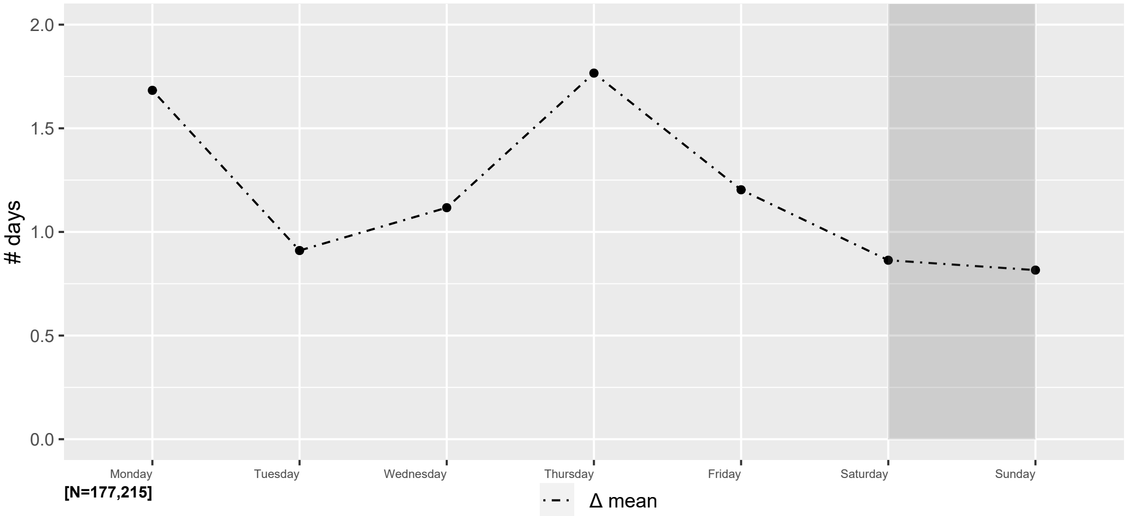
Source: IQTIG German hospital birth records for 2015-2016 restricted to the subsample of zero-precondition first-time mothers with pre-labor membrane ruptures (detailed in Table 3.A.2 and Table 3.A.3). Strata by maternal employment status. Own calculations.

Figure 3.A.5: Maternal Fitness & Births after Pre-Labor Membrane Ruptures Across Daily Hours



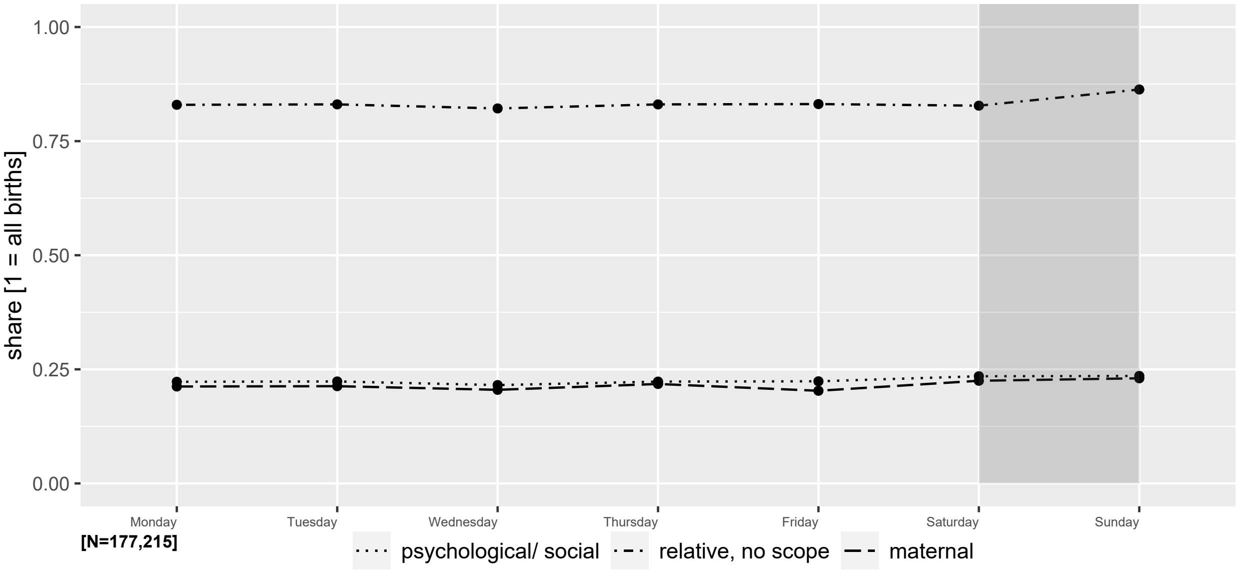
Source: IQTIG German hospital birth records for 2015-2016 restricted to the subsample of zero-precondition first-time mothers with pre-labor membrane ruptures (detailed in Table 3.A.2 and Table 3.A.3). Strata by maternal BMI. Own calculations.

Figure 3.A.6: Distribution Mean Due Date Prediction Error Across Weekdays



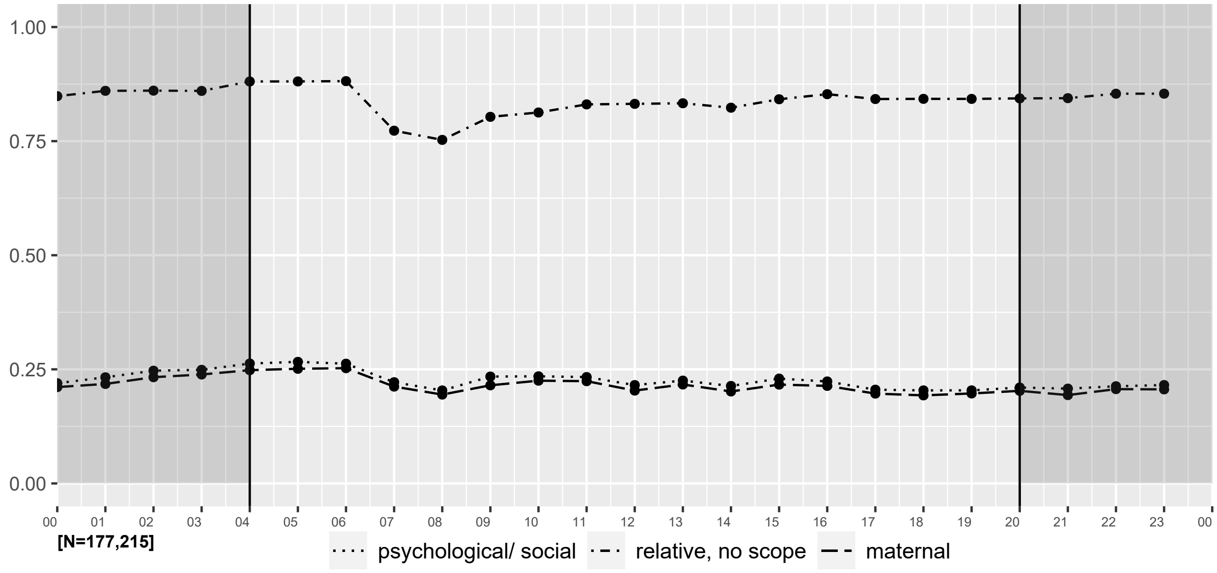
Source: IQTIG German hospital birth records for 2015-2016 restricted to the main analysis sample of zero-precondition first births (detailed in Table 3.A.2 and Table 3.A.3). $Birthdate - predictedduedate = \Delta$. Own calculations.

Figure 3.A.7: Distribution of Intervention Indications Across Weekdays



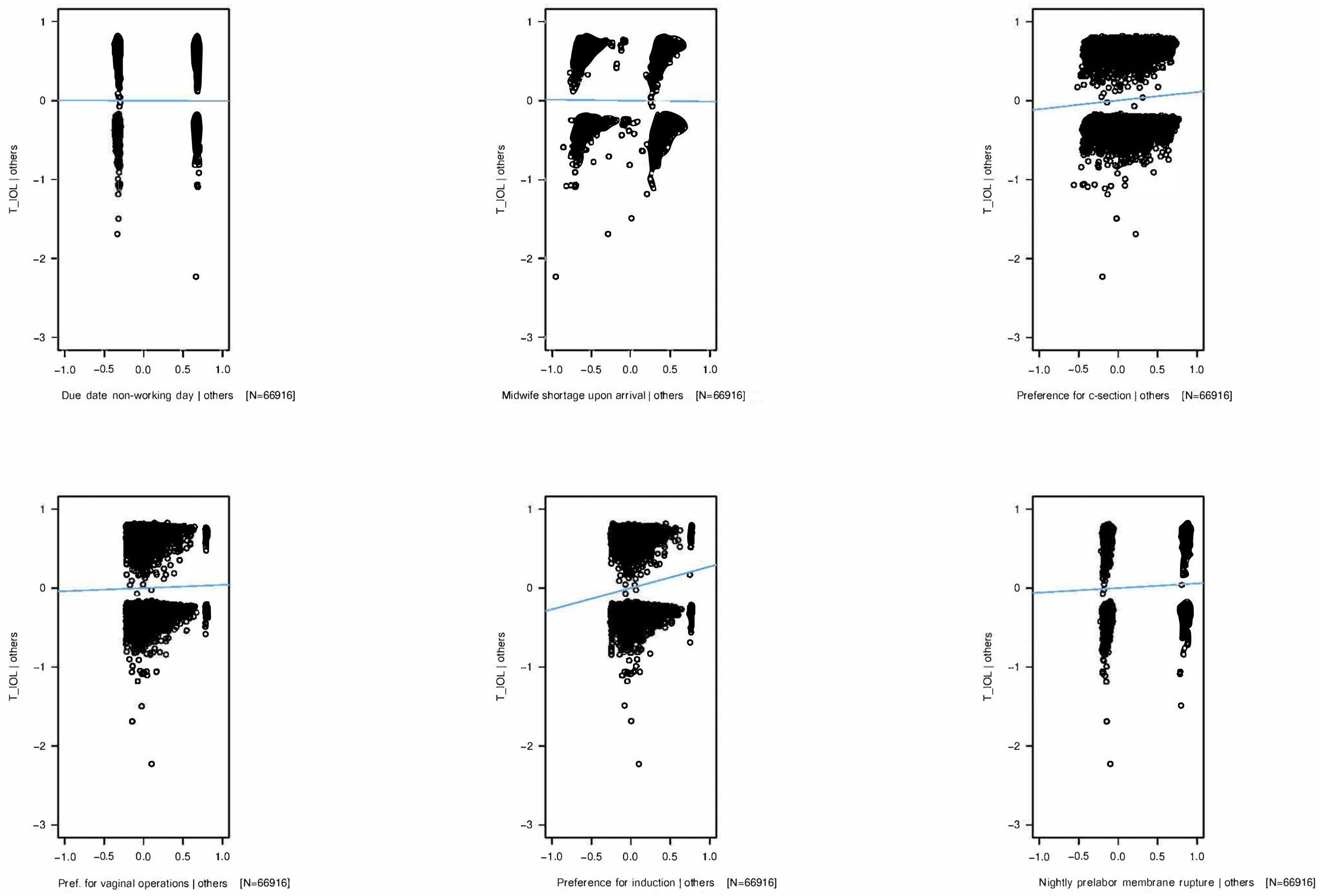
Source: IQTIG German hospital birth records for 2015-2016 restricted to the main analysis sample of zero-precondition first births (detailed in Table 3.A.2 and Table 3.A.3). Indications are grouped by implied medical decision scope for birth intervention, where *relative, no scope* comprise clearly stated medical conditions motivating (but not forcing) intervention. More vaguely defined are *psychological/social* conditions, and *maternal* refers to intervention on maternal request. Own calculations.

Figure 3.A.8: Distribution of Intervention Indications Across Daily Hours



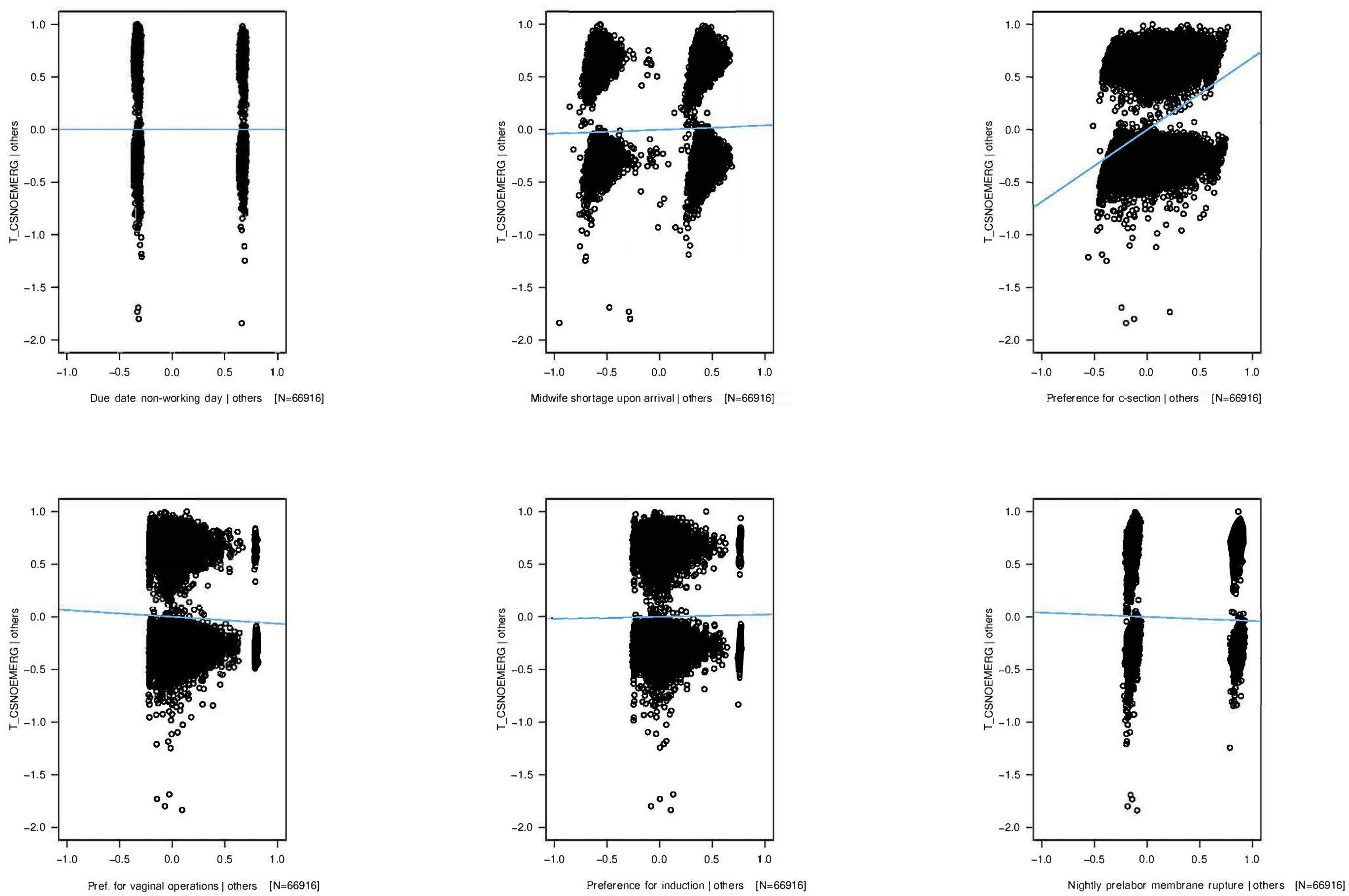
Source: IQTIG German hospital birth records for 2015-2016 restricted to the main analysis sample of zero-precondition first births (detailed in Table 3.A.2 and Table 3.A.3). Indications are grouped by implied medical decision scope for birth intervention, where *relative, no scope* comprise clearly stated medical conditions motivating (but not forcing) intervention. More vaguely defined are *psychological/social* conditions, and *maternal* refers to intervention on maternal request. Own calculations.

Figure 3.A.9: Added-Variable Plots of Induction & Instruments



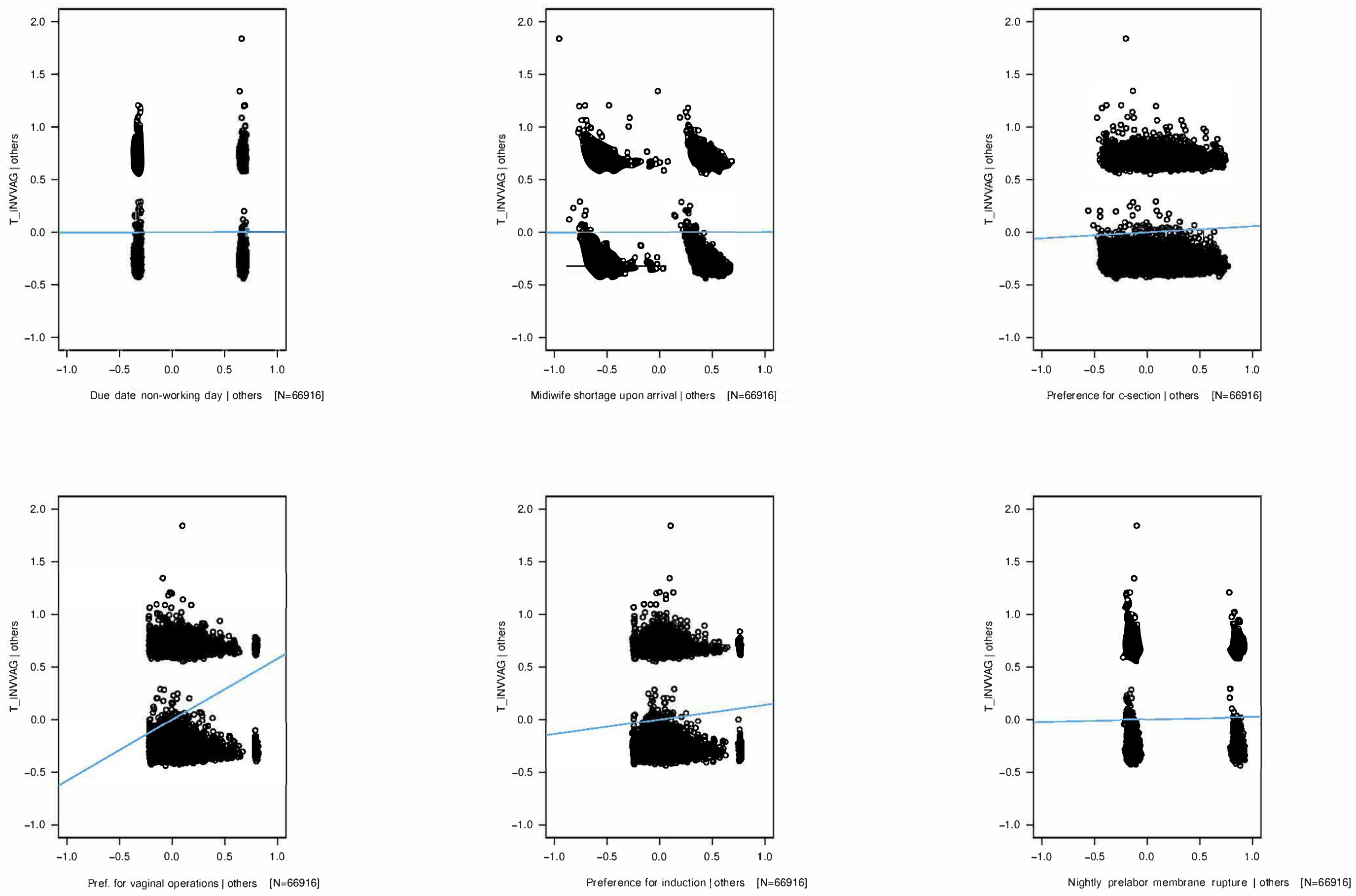
Source: IQTIG German hospital birth records, 2015-16, zero-precondition first births with information on obstetrician ids to compute obstetricians' preferences for intervention, see Table 3.A.2 and Table 3.A.3. Residual correlation conditional on core controls defined below Table 3.3. Own calculations.

Figure 3.A.10: Added-Variable Plots of Non-Emergency C-Section & Instruments



Source: IQTIG German hospital birth records, 2015-16, zero-precondition first births with information on obstetrician ids to compute obstetricians' preferences for intervention, see Table 3.A.2 and Table 3.A.3. Residual correlation conditional on core controls defined below Table 3.3. Own calculations.

Figure 3.A.11: Added-Variable Plots of Vaginally Invasive Procedures & Instruments



Source: IQTIG German hospital birth records, 2015-16, zero-precondition first births with information on obstetrician ids to compute obstetricians' preferences for intervention, see Table 3.A.2 and Table 3.A.3. Residual correlation conditional on core controls defined below Table 3.3. Own calculations.

Bibliography

Alberto Abadie and John F. Kennedy. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

Ran Abramitzky and Victor Lavy. How responsive is investment in schooling to changes in redistribution policies and in returns. *Econometrica*, 82(4):1241–1272, 2011.

ACOG. Clinical Guidance for Integration of the Findings of The ARRIVE Trial: Labor Induction Versus Expectant Management in Low-Risk Nulliparous Women. Technical report, American College of Obstetricians and Gynecologists, Washington DC, 2021.

Victoria M. Allen, Colleen M. O'Connell, and Thomas F. Baskett. Cumulative economic implications of initial method of delivery. *Obstetrics and Gynecology*, 108(3):549–555, 2006.

Douglas Almond, Kenneth Y. Chay, and David S. Lee. The Costs of Low Birth Weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

Isaiah Andrews and James H. Stock. Weak Instruments and What To Do About Them. Technical report, NBER Summer Institute, 2018.

Joshua D. Angrist. American education research changes tack. *Oxford Review of Economic Policy*, 20(2):198–212, 2004.

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

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press, Princeton, NJ., 2008.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion.* Number 8769. Princeton University Press, 1 edition, 2009.

Susan Athey and Guido W. Imbens. Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption. *NBER Working Paper*, 24963, 2018.

R. Axt-Fliedner, U. Wiegank, M. Friedrich, and K. Diedrich. Elektive Einleitung gegenüber spontanem Geburtsbeginn am Termin. *Gynäkologe*, 37(4):346–351, 2004.

Sascha O Becker and Ludger Woessmann. Was Weber wrong? A human capital theory of protestant economic history. *The Quarterly Journal of Economics*, 124(2):531–596, 2009.

Manudeep Bhuller, Gordon B Dahl, Katrine V Løken, and Magne Mogstad. Incarceration, Recidivism, and Employment. *Journal of Political Economy*, 128(4), 2020.

Carrie F. Bonsack, Anthony Lathrop, and Mary Blackburn. Induction of labor: Update and review. *Journal of Midwifery and Women's Health*, 59(6):606–615, 2014.

Alison Booth and Patrick Nolen. Choosing to compete: How different are girls and boys? *Journal of Economic Behavior and Organization*, 81(2):542–555, 2012.

Heather M. Bradford, Vicky Cárdenas, Katherine Camacho-Carr, and Mona T. Lydon-Rochelle. Accuracy of birth certificate and hospital discharge data: A certified nurse-midwife and physician comparison. *Maternal and Child Health Journal*, 11(6):540–548, 2007.

H. Shelton Brown. Physician demand for leisure: Implications for cesarean section rates. *Journal of Health Economics*, 15(2):233–242, 1996.

Alexandra Bruns. Das deutsche DRG-System: Die pauschale Geburt. *Deutsches Ärzteblatt*, 7(3):2014, 2014.

Alexandra Bruns. Analyse zum Änderungsvorschlag „Kostenstelle Kreißsaal". Technical report, Geburt e.V., Kronshagen, 2017.

Kasey Buckles and Melanie Guldi. Worth the Wait? The Effect of Early Term Birth on Maternal and Infant Health. *Journal of Policy Analysis and Management*, 36(4): 748–772, 2017.

BZgA. „Die Frauen können es, man lässt sie nur nicht!" Ein Gespräch mit Professor Alfred Rockenschaub über 50 Jahre Geburtshilfe und die wichtige Rolle der Frau. *FORUM Sexualaufklärung und Familienplanung*, 211(2):382–383, 2005.

A. Colin Cameron and Douglas L. Miller. A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, 50(2):317–372, 2015.

David Card, David Silver, and Alessandra Fenizia David. The Health Effects of Cesarean Delivery for low-risk first births. *NBER Working Paper Series*, 106(11):1323–1330, 2018.

Ursula Carle and Heinz Metzen. Literaturübersicht zum Stand derForschung, der praktischen Expertise und der pädagogischen Theorie. Eine wissenschaftliche Expertise des Grundschulverbandes. Frankfurtam Main: Grundschulverband (Wissenschaftliche Expertisen. In *Grundschulverband: Wissenschaftliche Expertisen*. Grundschulverband, Frankfurt am Main, 2014.

Suzan L. Carmichael and Jonathan M. Snowden. The ARRIVE Trial – Interpretation from an Epidemiologic Perspective. *Journal of Midwifery & Women's Health.*, 64(5):657–663, 2019.

Scott E. Carrell, Bruce I. Sacerdote, and James E. West. From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation. *Econometrica*, 81(3):855–882, 2013.

Sarah Carter, Amos Channon, and Ann Berrington. Socioeconomic risk factors for labour induction in the United Kingdom. *BMC Pregnancy and Childbirth*, 20(146):284–300, 2020.

Daniele Checchi and Maria De Paola. The effect of multigrade classes on cognitive and non- cognitive skills. Causal evidence exploiting minimum class size rules in Italy. *Economics of Education Review*, 67:235–253, 2018a. URL `https://www.sciencedirect.com/science/article/abs/pii/S0272775718300943`.

Daniele Checchi and Maria De Paola. The effect of multigrade classes on cognitive and non-cognitive skills. Causal evidence exploiting minimum class size rules in Italy. *Economics of Education Review*, 67(1):235–253, 2018b.

Victor Chernozhukov, Christian Hansen, and Michael Jansson. Admissible Invariant Similar Tests for Instrumental Variables Regression. *Econometric Theory*, 25(3):806–818, 2009.

Raj Chetty, John N. Friedman, and Jonah E. Rockoff. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 104(9):2633–2679, 2014a.

Raj Chetty, Nathaniel Hendren, Patrick Kline, Emmanuel Saez, and Nicholas Turner. Is the United States still a land of opportunity? Recent trends in intergenerational mobility. *American Economic Review*, 104(5):141–147, 2014b.

Damon Clark and Emilia Del Bono. The Long-Run Effects of Attending an Elite School: Evidence from the United Kingdom. *American Economic Journal: Applied Economics*, 8(1):150–176, 2016.

Dominiek Coatesid, Angela Makris, Christine Catling, Amanda Henry, Vanessa Scarf, Nicole Watts, Deborah Fox, Purshaiyna Thirukumar, Vincent Wong, Hamish Russell, and Caroline Homer. A systematic scoping review of clinical indications for induction of labour. *Plos One*, 15(1):1–11, 2020.

James S. Coleman. Equality of Educational Opportunity Study (EEOS). *Equity and Excellence in Education*, 6(5):1–45, 1968.

Kalena E. Cortes and Joshua S. Goodman. Ability-tracking, instructional time, and better pedagogy: The effect of double-dose Algebra on student achievement. *The American Economic Review*, 104(5):400–405, 2014.

Ana Costa-Ramón, Mika Kortelainen, Ana Rodríguez-González, and Lauri Sääksvuori. The Long-Run Effects of Cesarean Sections. *VATT Working Papers*, 125(1):1–48, 2019.

Ana María Costa-Ramón, Ana Rodríguez-González, Miquel Serra-Burriel, and Carlos Campillo-Artero. It's about time: Cesarean sections and neonatal health. *Journal of Health Economics*, 59(1):46–59, 2018.

Julie Berry Cullen, Brian A. Jacob, and Steven Levitt. The effect of school choice on participants: Evidence from randomized lotteries. *Econometrica*, 74(5):1191–1230, 2006.

Janet Currie. Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and human capital development. *Journal of Economic Literature*, 47(1):87–122, 2009.

David J Deming, Justine S Hastings, Thomas J Kane, and Douglas O Staiger. School Choice, School Quality, and Postsecondary Attainment. *The American Economic Review*, 104(3):991–1013, 2014.

DGGG. Empfehlungen zum Vorgehen beim vorzeitigen Blasensprung S1 Leitlinie. Technical report, Deutsche Gesellschaft für Gynäkologie und Geburtshilfe, Göttingen, 2006.

DGGG. Geburtseinleitung Leitlinienklasse S2k Stand Dezember 2020. Technical report, Deutsche Gesellschaft für Gynäkologie und Geburtshilfe, Göttingen, 2020a.

DGGG. Sectio caesarea Leitlinienklasse S3 Stand Juni 2020. Technical report, Deutsche Gesellschaft für Gynäkologie und Geburtshilfe, Göttingen, 2020b.

DGGG. Leitlinie zum Management von Dammrissen III. und IV. Grades nach vaginaler Geburt S2k Stand Dezember 2020. Technical report, Deutsche Gesellschaft für Gynäkologie und Geburtshilfe, Göttingen, 2020c.

DHV and DGGG. DHV und DGGG fordern Korrekturen am Gesetzentwurf für die 1:1-Betreuung in der klinischen Geburtshilfe. Technical report, Deutscher Hebammen Verband und Deutsche Gesellschaft für Gynäkologie und Geburtshilfe, Berlin, 2020.

David Dranove and Paul Wehner. Physician-induced demand for childbirths. *Journal of Health Economics*, 13(1):61–73, 1994.

Esther Duflo, Pascaline Dupas, and Michael Kremer. Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *The American Economic Review*, 101(5):1739–1774, 2011.

EENEE. The Impact of School Size and Consolidations on Quality and Equity in Education. Technical Report 19, EENEE (European Expert Network on Economics of Education), 2015.

A. Feige. Budget und organisationsaspekte einer geburtshilflichen abteilung. *Gynäkologe*, 41(1):42–48, 2008.

David N. Figlio and Joe A. Stone. Are private schools really better? *Research in Labor Economics*, 18(1):115–140, 1999.

Joshua S. Gans and Andrew Leigh. Born on the first of July: An (un)natural experiment in birth timing. *Journal of Public Economics*, 1(1):1–18, 2008.

Susan Garthus-Niegel, Cecilie Knoph, Tilmann von Soest, Christopher S. Nielsen, and Malin Eberhard-Gran. The Role of Labor Pain and Overall Birth Experience in the Development of Posttraumatic Stress Symptoms: A Longitudinal Cohort Study. *Birth*, 41(1):108–115, 2014.

Ilka Gerhardts. The Economics of Labor & Patients' Health Outcomes: Evidence from Childbirth in Germany. *mimeo, University of Munich (LMU)*, 2024.

Ilka Gerhardts, Uwe Sunde, and Larissa Zierow. Effects of Multi-grade Classes in Primary Schools on Educational. *mimeo, University of Munich (LMU)*, 2021a.

Ilka Gerhardts, Uwe Sunde, and Larissa Zierow. Class Composition and Educational Outcomes: Evidence from the Abolition of Denominational Schools. *mimeo, University of Munich (LMU)*, 2021b.

Salvatore Gizzo, Carlo Saccardi, Tito Silvio Patrelli, Stefania Di Gangi, Elisa Breda, Simone Fagherazzi, Marco Noventa, Donato D'antona, and Giovanni Battista Nardelli. Fertility rate and subsequent pregnancy outcomes after conservative surgical techniques in postpartum hemorrhage: 15 years of literature. *Fertility and Sterility*, 99(1):2097–2107, 2013.

Paul Glewwe, Michael Kremer, and Sylvie Moulin. Many Children Left Behind? Textbooks and Test Scores in Kenya. *American Economic Journal: Applied Economics*, 1(1):112–135, 2009.

Uri Gneezy, Muriel Niederle, Aldo Rustichini, Dan Brodkey, Stefano Della Vigna, Gerhard Orosel, Nikita Piankov, Al Roth, and Lise Vesterlund. Performance in competitive Environments: Gender differences. *The Quarterly Journal of Economics*, 118(3):1049–1074, 2003.

GNPI. Betreuung von Neugeborenen in der Geburtsklinik - S2k Leitlinie. *AWMF online*, 024-005:1–25, 2022.

Henci Goer. Parsing the ARRIVE Trial: Should First-Time Parents Be Routinely Induced at 39 Weeks? Technical report, Lamaze International, Washington DC, 2018.

Karin Gottvall and Ulla Waldenström. Does a traumatic birth experience have an impact on future reproduction? *BJOG: An International Journal of Obstetrics and Gynaecology*, 109(3):254–260, 2002.

Jeremy Greenwood, Nezih Guner, Georgi Kocharkov, and Cezar Santos. Technology and the Changing Family: A Unified Model of Marriage, Divorce, Educational Attainment, and Married Female Labor-Force Participation. *American Economic Journal: Macroeconomics*, 8(1):1–41, 2016.

William A. Grobman, Madeline M. Rice, Uma M. Reddy, Alan T.N. Tita, Robert M. Silver, Gail Mallett, Kim Hill, Elizabeth A. Thom, Yasser Y. El-Sayed, Annette Perez-Delboy, Dwight J. Rouse, M.D. George R. Saade, Kim A. Boggess, Suneet P. Chauhan, Jay D. Iams, Edward K. Chien, Brian M. Casey, Ronald S. Gibbs, Sindhu K. Srinivas, Geeta K. Swamy, Hyagriv N. Simhan, and George A. Macones. Labor Induction versus Expectant Management in Low-Risk Nulliparous Women. *The NEW ENGLAND JOURNAL of MEDICINE*, 379(6):513–523, 2018.

İsmet Gün, Bülent Doğan, and Özkan Özdamar. Long- and short-term complications of episiotomy. *Turkish Journal of Obstetrics and Gynecology*, 13(2):144–148, 2016.

Martin Halla, Harald Mayr, Gerald J. Pruckner, and Pilar García-Gómez. Cutting fertility? Effects of cesarean deliveries on subsequent fertility and maternal labor supply. *Journal of Health Economics*, 72(1):102–125, 2020.

Eric Hanushek, John F. Kain, Jacob M. Markman, and Steven G. Rivkin. Does peer ability affect student achievement? *Journal of Applied Econometrics*, 18(5):527–544, 2003.

Lorie M. Harper, Aaron B. Caughey, Anthony O. Odibo, Kimberly A. Roehl, Qiuhong Zhao, and Alison G. Cahill. Normal progress of induced labor. *Obstetrics and Gynecology*, 321(3):25–31, 2012.

John A. Hattie. Classroom composition and peer effects. *International Journal of Educational Research*, 37(5):449–481, 2002.

James Heckman, Lance Lochner, and Christopher Taber. General-Equilibrium Treatment Effects: A Study of Tuition Policy. *The American Economic Review*, 88(2):381–386, 1998.

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

James J. Heckman. The Economics of Inequality: The Value of Early Childhood Education. *American Educator*, 1(1):31–47, 2011.

Friederike Heinzel and Katja Koch. *Individualisierung im Grundschulunterricht*. Springer, Wiesbaden, 2017.

Marcel Helbig and Rita Nikolai. Die Unvergleichbaren. Der Wandel der Schulsysteme in den deutschen Bundesländern seit 1949. Technical report, Klinkhardt, Bad Heilbrunn, 2015.

Andrew J. Hill. The girl next door: The effect of opposite gender friends on high school achievement. *American Economic Journal: Applied Economics*, 7(3):147–177, 2015.

InEK. aG-DRG-Version und Pflegeerlöskatalog 2021. Technical report, InEK Institut für das Entgeltsystem im Krankenhaus, Siegburg, 2021.

IQTIG. Qualitätsreport 2016. Technical report, Institut für Qualitätssicherung und Transparenz im Gesundheitswesen (i.A. Gemeinsamer Bundesausschuss), Berlin, 2017.

Jon M Iversen and Hans Bonesrønning. Conditional gender peer effects? *Journal of Behavioral and Experimental Economics*, 55:19–28, 2015.

Christine Jachetta. *Topics in Health Economics*. PhD thesis, University of Illinois at Urbana-Champaign, 2016.

Anne Katz Jacobson. Are we losing the art of midwifery? *Journal of Nurse-Midwifery*, 38 (3):168–169, 1993.

Mireille Jacobson, Maria Kogelnik, and Heather Royer. Holiday, Just One Day Out of Life: Birth Timing and Post-Natal Outcomes. *NBER Working Paper Series*, 27236, 2020.

Wonjeong Jeong, Sung In Jang, Eun Cheol Park, and Jin Young Nam. The effect of socioeconomic status on all-cause maternal mortality: A nationwide population-based cohort study. *International Journal of Environmental Research and Public Health*, 17 (12):1–13, 2020.

John Jolly, James Walker, and Kalvinder Bhabra. Subsequent obstetric performance related to primary mode of delivery. *BJOG: An International Journal of Obstetrics and Gynaecology*, 106(3):227–232, 1999.

Sam Jones. Class size versus class composition: What matters for learning in East Africa? *WIDER Working Paper Series*, 065, 2013.

Hendrik Jürges. Financial incentives, timing of births, and infant health: a closer look into the delivery room. *European Journal of Health Economics*, 18(1):195–208, 2017.

Hendrik Jürges and Juliane Köberlein. What explains DRG upcoding in neonatology? The roles of financial incentives and infant health. *Journal of Health Economics*, 43(2): 13–26, 2015.

KBV. Einheitlicher Bewertungsmaßstab (EBM). 01:1–1779, 2020.

Petra Kolip, Hans-Dieter Nolting, and Karsten Zich. Kaiserschnittgeburten – Entwicklung und regionale Verteilung. Technical report, Bertelsmann Stiftung, Gütersloh, 2012.

Alan B Krueger. Experimental estimates of education production functions. *The Quarterly Journal of Economics*, 114(2):497–532, 1999.

Victor Lavy and Analía Schlosser. Mechanisms and impacts of gender peer effects at school. *American Economic Journal: Applied Economics*, 3(2):1–33, 2011.

Victor Lavy, M. Daniele Paserman, and Analia Schlosser. Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom. *Economic Journal*, 122(559):208–237, 2012.

Soohyung Lee, Lesley J. Turner, Seokjin Woo, and Kyunghee Kim. The impact of school and classroom gender composition on educational achievement. *NBER Working Paper Series*, 113(4):1–40, 2015.

Edwin Leuven and Marte Rønning. Classroom Grade Composition and Pupil Achievement. *Economic Journal*, 126(593):1164–1192, 2016.

Lisa D. Levine, Katheryne L. Downes, Michal A. Elovitz, Samuel Parry, Mary D. Sammel, and Sindhu K. Srinivas. Mechanical and Pharmacologic Methods of Labor Induction: A Randomized Controlled Trial. *Obstetrics & Gynecology*, 128(6):1357–1364, 2016.

Karl Lichtblau. 50 Jahre Saarland: Wirtschaft Saarland 1959 bis 2009 Wie hat sich das Saarland in den letzten 50 Jahren entwickelt - ein Bundesländervergleich. Technical report, Institut der deutschen Wirtschaft Köln (IW Consult GmbH), Köln, 2009.

Elly-Ann Lindström and Erica Lindahl. The Effect of Mixed-Age Classes in Sweden. *Scandinavian Journal of Educational Research*, 55(2):121–144, 2011.

Angela W. Little. Multigrade teaching: towards an international research and policy agenda. *International Journal of Educational Development*, 21(6):481–497, 2001.

Angela W Little. Learning and teaching in multigrade settings. *Paper commissioned for the EFA Global Monitoring Report 2005, The Quality Imperative*, 2004.

Petter Lundborg, Anton Nilsson, and Dan-Olof Rooth. Parental Education and Offspring Outcomes: Evidence from the Swedish Compulsory Schooling Reform. *IZA Discussion Paper Series*, 6570, 2012.

Ulrike Lutz and Petra Kolip. Die GEK-Kaiserschnittstudie. In GEK-Gmünder-Ersatzkasse, editor, *Schriftenreihe zur Gesundheitsanalyse.* Asgard Verlag, Bremen/ Schwäbisch Gmünd, 2006.

Mona Lydon-Rochelle, Victoria L. Holt, Diane P. Martin, and Thomas R. Easterling. Association Between Method of Delivery and Maternal Rehospitalization. *JAMA*, 283 (18):2411–2416, 2000.

John Lynch, Aurélie Meunier, Rhiannon Pilkington, and Stefanie Schurer. Baby Bonuses and Early-Life Health Outcomes: Using Regression Discontinuity to Evaluate the Causal Impact of an Unconditional Cash Transfer. *IZA Discussion Papers*, 12230(1):1–54, 2019.

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

DeWayne A. Mason and Robert B. Burns. 'Simply No Worse and Simply No Better' May Simply Be Wrong: A Critique of Veenman's Conclusion About Multigrade Classes. *Review of Educational Research*, 66(3):307–322, 1996.

Patrick J. McEwan. Evaluating multigrade school reform in Latin America. *Comparative Education*, 44(4):465–483, 2008.

Ekaterina Mishanina, Ewelina Rogozinska, Tej Thatthi, Rehan Uddin-Khan, Khalid S. Khan, and Catherine Meads. Use of labour induction and risk of cesarean delivery: a systematic review and meta-analysis. *CMAJ*, 186(9):665–673, 2014.

Model Professional Code for Physicians in Germany. MBO-Ä 1997-**The Resolutions of the 121st German Medical Assembly 2018 in Erfurt as amended by a Resolution of the Executive Board of the German Medical Association on 14/12/2018, I. Principles Art.1 (1), 1997.

Aidan Mulkeen and Cathal Higgings. Multigrade Teaching in Sub-Saharan Africa. *World Bank Working Paper Series*, 173, 2009.

Ioannis Mylonas and Klaus Friese. The indications for and risks of elective cesarean section. *Deutsches Ärzteblatt International*, 112(29-30):489–495, 2015.

Karen Norberg and Juan Pantano. Cesarean sections and subsequent fertility. *Journal of Population Economics*, 29(1):5–37, 2016.

Charles O'Donovan and James O'Donovan. Why do women request an elective cesarean delivery for non-medical reasons? A systematic review of the qualitative literature. *Birth*, 45(2):109–119, 2018.

Vicky O'Dwyer, Sarah O'Kelly, Bernadette Monaghan, Ann Rowan, Nadine Farah, and Michael J. Turner. Maternal obesity and induction of labor. *Acta Obstetricia et Gynecologica Scandinavica*, 92(12):1414–1418, 2013.

Man Wah Pang, Tse Ngong Leung, Tze Kin Lau, and Tony Kwok Hang Chung. Impact of first childbirth on changes in women's preference for mode of delivery: Follow-up of a longitudinal observational study. *Birth*, 35(2):121–128, 2008.

Petra Persson and Maya Rossin-Slater. Family ruptures, stress, and the mental health of the next generation. *American Economic Review*, 108(4-5):1214–1252, 2018.

Jörn-Steffen Pischke and Till Von Wachter. Zero returns to compulsory schooling in Germany: Evidence and interpretation. *The Review of Economics and Statistics*, 90(3): 592–598, 2005.

Holly Priddis, Hannah G. Dahlen, Virginia Schmied, Annie Sneddon, Christine Kettle, Chris Brown, and Charlene Thornton. Risk of Recurrence, Subsequent Mode of Birth and Morbidity for Women who Experienced Severe Perineal Trauma in a First Birth in New South Wales between 2000–2008: A Population Based Data Linkage Study. *BMC Pregnancy and Childbirth*, 89(13):1471–1493, 2013.

David Roodman, Morten O. Nielsen, James G. MacKinnon, and Matthew D Webb. Fast and wild: Bootstrap inference in Stata using boottest. *Stata Journal*, 19(1):4–60, 2019.

Jesse M Rothstein. Good principals or good peers? Parental valuation of school characteristics, tiebout equilibrium, and the incentive effects of competition among jurisdictions. *The American Economic Review*, 96(4):1333–1350, 2006.

Sandra Ilona Rummel. *Kosten und Erlöse bei der Abrechnung geburtshilflicher Leistungen nach dem System der Diagnosis-Related-Groups (DRG).* PhD thesis, LMU Munich, 2007.

Jean V. Russell, Kenneth J. Rowe, and Peter W. Hill. Effects of Multigrade Classes on Student Progress in Literacy and Numeracy: Quantitative Evidence and Perceptions of Teachers and School Leaders. In *Annual Meeting of the Australian Association for Research in Education*, pages 212–243, Adelaide, 1998.

Elsa Lena Ryding, Mirjam Lukasse, Hildur Kristjansdottir, Thora Steingrimsdottir, Berit Schei, Ann Tabor, Helle Karro, An Sofie Van Parys, Anne Mette Schroll, Made Laanpere, and Anne Marie Wangel. Pregnant women's preference for cesarean section and subsequent mode of birth – a six-country cohort study. *Journal of Psychosomatic Obstetrics and Gynecology*, 37(3):75–83, 2016.

Eleanor Sanderson and Frank Windmeijer. A weak instrument F-test in linear IV models with multiple endogenous variables. *Journal of Econometrics*, 190(2):212–221, 2016.

Anton Scharl, Frank Louwen, and Christian Albring. Offener Brief anlässlich der Öffentlichen Anhörung im Ausschuss für Gesundheit des Deutschen Bundestages am Mittwoch, 26. Juni 2019, zum Thema „Entwurf eines Gesetzes zur Reform der Hebammenausbildung“. Technical report, German Board and College of Obstetrics and Gynecology, Berlin, 2019.

Lisa Schulkind and Teny Maghakian Shapiro. What a difference a day makes: Quantifying the effects of birth timing manipulation on infant health. *Journal of Health Economics*, 33(1):139–158, 2014a.

Lisa Schulkind and Teny Maghakian Shapiro. What a difference a day makes: Quantifying the effects of birth timing manipulation on infant health. *Journal of Health Economics*, 33(1):139–158, 2014b.

Christiane Schwarz, Rainhild Schäfers, Christine Loytved, Peter Heusser, Michael Abou-Dakn, Thomas König, and Bettina Berger. Temporal trends in fetal mortality at and beyond term and induction of labor in Germany 2005–2012: data from German routine perinatal monitoring. *Archives of Gynecology and Obstetrics*, 293(2):335–343, 2016.

Clarissa M Schwarz. *Entwicklung der geburtshilflichen Versorgung – am Beispiel geburtshilflicher Interventionsraten 1984-1999 in Niedersachsen.* PhD thesis, Technischen Universität Berlin, 2008.

David Sims. A Strategic Response to Class Size Reduction: Combination Classes and Student Achievement in California. *Journal of Policy Analysis and Management*, 29(3): 451–478, 2010.

Andrei Smarandache, Theresa H.M. Kim, Yvonne Bohr, and Hala Tamim. Predictors of a negative labour and birth experience based on a national survey of Canadian women. *BMC Pregnancy and Childbirth*, 16(1):903–920, 2016.

Enrico Spolaore and Romain Wacziarg. How Deep Are the Roots of Economic Development? *Journal of Economic Literature*, 51(2):232–347, 2013.

Melvin Stephens and Dou-Yan Yang. Compulsory Education and the Benefits of Schooling. *American Economic Review*, 104(6):1777–1792, 2014.

Hege Therese Størksen, Susan Garthus-Niegel, Samantha S. Adams, Siri Vangen, and Malin Eberhard-Gran. Fear of childbirth and elective caesarean section: A population-based study. *BMC Pregnancy and Childbirth*, 15(1):1–11, 2015.

A Tammaa, W Umek, and M. Wunderlich. Leitlinie zum Management von Dammrissen III. und IV. Grades nach vaginaler Geburt. *Speculum - Zeitschrift für Gynäkologie und Geburtshilfe*, 25(3):15–15, 2007.

Jaime Thomas. Combination classes and educational achievement. *Economics of Education Review*, 31(6):1058–1066, 2012.

Ioannis Tsakiridis, Apostolos Mamopoulos, Apostolos Athanasiadis, and Themistoklis Dagklis. Induction of Labor: An Overview of Guidelines. *Obstetrical and Gynecological Survey*, 75(1):61–72, 2020.

Sibil Tschudin, Judith Alder, Stephanie Hendriksen, Johannes Bitzer, Karoline Aebi Popp, Rosanna Zanetti, Irene Hösli, Wolfgang Holzgreve, and Verena Geissbühler. Previous birth experience and birth anxiety: Predictors of caesarean section on demand? *Journal of Psychosomatic Obstetrics and Gynecology*, 30(3):175–180, 2009.

Peter Tyson. The Hippocratic Oath Today. *NOVA*, 2001.

Anjel Vahratian, Jun Zhang, James F. Troendle, Anthony C. Sciscione, and Matthew K. Hoffman. Labor progression and risk of cesarean delivery in electively induced nulliparas. *Obstetrics and Gynecology*, 105(4):698–704, 2005.

Simon Veenman. Cognitive and Noncognitive Effects of Multigrade and Multi-Age Classes: A Best-Evidence Synthesis. *Review of Educational Research*, 65(4):319–381, 1995.

Eva Vivalt. Heterogeneous Treatment Effects in Impact Evaluation. *The American Economic Review*, 105(5):467–470, 2015.

Francis P.J.M. Vrouenraets, Frans J.M.E. Roumen, Cary J.G. Dehing, Eline S.A. Van Den Akker, Maureen J.B. Aarts, and Esther J.T. Scheve. Bishop score and risk of cesarean delivery after induction of labor in nulliparous women. *Obstetrics and Gynecology*, 105 (4):690–697, 2005.

Matthea Wagener. *Gegenseitiges Helfen. Soziales Lernen im jahrgangsgemischten Unterricht.* Springer, Wiesbaden, 2014.

Diane Whitmore. Resource and peer impacts on girls' academic achievement: Evidence from a randomized experiment. *The American Economic Review*, 95(2):199–203, 2005.

WHO. WHO Statement on Caesarean Section Rates. Technical report, World Health Organisation, Geneva, 2015.

WHO. Recommendations: Induction of Labour at or beyond Term. Technical report, World Health Organisation, Geneva, 2018.

Katherine B. Wolfe, Rocco A. Rossi, and Carri R. Warshak. The effect of maternal obesity on the rate of failed induction of labor. *American Journal of Obstetrics and Gynecology*, 205(2):128.e1–128.e7, 2011.

Andrea Mary Woolner, Dolapo Ayansina, Mairead Black, and Sohinee Bhattacharya. The Impact of Third- or Fourth-Degree Perineal Tears on the Second Pregnancy: A cohort study of 182,445 Scottish Women. *PLoS One*, 14(4):1–18, 2019.

C. M. Zahn and E. R. Yeomans. Postpartum hemorrhage: Placenta accreta, uterine inversion, and puerperal hematomas. *Clinical Obstetrics and Gynecology*, 33(3):422–431, 1990.