Child Penalty Estimation and Mothers' Age at First Birth

Valentina Melentyeva*   Lukas Riedel†

December 12, 2025‡

We show that the widespread approach to estimate the career costs of motherhood – so-called "child penalties" – is prone to produce biased results, as it pools first-time mothers of all ages without accounting for their differences in characteristics and outcomes. We propose a novel method building on the recent advances in the difference-in-differences literature to address this issue. Applied to German administrative data, our method yields 28 percent larger post-birth earnings losses than the conventional approach. We document meaningful effect heterogeneity by maternal age in both magnitude and interpretation, highlighting its key role in understanding the impact of motherhood.

Keywords: child penalty, maternal labor supply, heterogeneous treatment effects, event study

JEL: J13, J16, J31, C23

* Tilburg University and RFBerlin (v.melentyeva@tilburguniversity.edu)

† ZEW Mannheim (lukas.riedel@zew.de)

‡ We are very grateful to Jérôme Adda, James Banks, Anna Bindler, Richard Blundell, Monica Costa Dias, Clément de Chaisemartin, Scott Cunningham, Thomas Dohmen, Christian Dustmann, Sergei Guriev, Lena Janys, Xavier Jaravel, Hyejin Ku, Petter Lundborg, Barbara Petrongolo, Pia Pinger, Erik Plug, Uta Schönberg, Sebastian Siegloch, Bettina Siflinger, Holger Stichnoth, Tom Zimmermann, and seminar and conference participants at the University College London, University of Cologne, University of Bonn, University of Manchester, ZEW – Leibniz Centre for European Economic Research, IZA – Institute of Labor Economics, Lund University, Berlin Applied Micro Seminar, EEA Annual Meeting 2023, AFSE Annual Congress 2023, EALE Annual Meeting 2023, CReAM-RFBerlin Workshop 2023, ESEM Congress 2024, and Tinbergen Institute – Labor Seminar Series 2025 for helpful discussions and suggestions.

I Introduction

The past century has been marked by major successes of the women's rights movement. Especially in developed countries, women gained – among other achievements – widespread and unrestricted access to education and employment. Nevertheless, gender inequality remains prevalent in labor markets around the world, and one particular reason persists to this day: motherhood is still costly for the careers of women. A large body of work documents that gender differences in the costs of parenthood are the major driver of the remaining gender inequality in labor market outcomes (see the reviews by Blau and Kahn 2017; Petrongolo and Ronchi 2020; Andrew et al. 2021). Therefore, correctly tracking the dynamics of the career costs of motherhood and studying the underlying mechanisms is crucial for understanding gender inequality and giving informed policy advice. In this paper, we show that the most popular approach to estimating the labor market impact of motherhood is likely to produce biased results, as it does not account for differences in outcomes and characteristics of mothers by their age at first childbirth. We propose a solution that addresses these biases and demonstrate that considering heterogeneity is not only methodologically important but also offers an opportunity to gain a deeper understanding of the impact of motherhood on women's labor market outcomes.

The career costs of motherhood have been both subject of research and a recurring topic in public debates for a long time. Recently, an approach based on event studies around the first childbirth – estimating so called "child penalties" – received widespread attention as it provides a straightforward and intuitive way to visualize the career impact of childbirth. The paper by Kleven, Landais, and Søgaard (2019) popularized the method and gained more than 1,900 citations over six years. The authors crucially contributed to an understanding of the challenges of combining motherhood and a career among a wide range of audiences.

Researchers have been actively using this event study approach to estimate the labor market effects of motherhood, to measure gender inequality within and across countries, and to evaluate family-related policies. However, these conventional event studies – linear regression models with event-time indicators and covariates – pool together younger and older first-time mothers to estimate average effects of motherhood. They implicitly assume that mothers of different ages are comparable and that the effects of childbirth are uniform for them. We demonstrate that these assumptions are unlikely to hold. First, mothers are very different depending on their age at birth and the effects of motherhood vary across women who give birth at different stages of life and career development. As the literature shows and as we systematically review, the age at which a woman gives birth to her first child is strongly correlated with both pre- and post-birth outcomes as well as human capital, her family's socio-economic status and other characteristics that are relevant in the labor market. Second, the impact of motherhood is prone to change over time after birth as mothers adjust their labor supply in response to changing childcare duties and constraints set by policies.

This heterogeneity by age at birth is not a problem per se. In general, heterogeneous characteristics or effects across certain groups is not an issue as long as the result of an estimation still represents a correctly weighted average of group-specific effects. However, age at birth is not just a characteristic, such as education or occupation, but also a dimension of *treatment timing*. Recent literature on difference-in-differences (DiD) models with staggered treatment adoption and heterogeneous treatment effects (see the summaries by Roth et al. 2023; de Chaisemartin and D'Haultfoeuille 2023) demonstrates that effect heterogeneity by treatment timing causes an event study design to produce something very different than a correct average effect over groups. Applying the insights from the literature (Goodman-Bacon 2021; Sun and Abraham 2021), we show that conventional event studies around childbirth are an illustration of this problem, as they are prone to produce substantially biased estimates of the effects of motherhood due to the pronounced heterogeneity by age at birth. We demonstrate that there are two underlying issues, typically coined as forbidden comparisons and contamination, that both arise in settings where multiple groups have heterogeneous treatment effects and are treated at different points in time. Forbidden comparisons are made because an event study cannot properly align relative time around treatment with the time dimension that determines effect heterogeneity, which leads to already-treated units entering the control group. In the context of childbirth, forbidden comparisons mean that mothers who have already given birth end up as part of the control group. Contamination happens since an event study contains a set of event time dummies such that the estimate of the average treatment effect for each relative time period is conditional on the average effects from all other periods. Since one average is not sufficient to account for heterogeneous effects across several groups, each estimate for a relative time period can be contaminated by the treatment effects from all other periods. With childbirth, contamination means that estimates for post-birth periods can include effects from periods prior to birth and vice versa. The consequence of both problems for conventional event studies around the first childbirth is that estimates are likely to be biased and that pre-trends might not be informative about the plausibility of the key identifying assumption of parallel trends. We illustrate how both issues materialize and lead to substantial biases when estimating the effects of motherhood using conventional event studies.

The problems caused by heterogeneous treatment effects in settings with staggered adoption have been addressed by newly developed approaches that allow for heterogeneous effects and include only non-treated units in the control groups (see, among others, Cengiz et al. 2019; de Chaisemartin and D'Haultfoeuille 2020; Sun and Abraham 2021; Callaway and Sant'Anna 2021; Borusyak, Jaravel, and Spiess 2024). However, the key decision in an event study design is the choice of a control group that satisfies the identifying assumption of parallel trends – that in the post-treatment period, the control and treated groups would have followed the same outcome trajectories in the absence of treatment. In the case of childbirth, the potentially available control groups are mothers who have not-yet given birth, childless women, and men. Given the systematic differences between mothers and non-mothers as well as men, the differential experiences of women and men in the labor market and the selection into having children, the validity of the latter two as control groups hinges on assumptions that are unlikely to hold. The heterogeneity of mothers by their age at birth, in turn, implies that not all of the not-yet treated mothers can serve as a suitable control group, as much older mothers are not comparable to younger ones.

Therefore, we propose to use only women with close ages at birth as control units. For each group of mothers who give birth at a given age, we choose a control group exclusively from the pre-birth observations of not-yet-treated mothers who give birth at slightly older ages. With this approach, we use the strong correlation between age at first birth and labor market outcomes to our advantage and bring together the most comparable mothers in the treatment and control groups. We build on the DiD models under staggered adoption and heterogeneous effects proposed by Sun and Abraham (2021), Callaway and Sant'Anna (2021), and Wing, Freedman, and Hollingsworth (2024) and extend the potential outcomes model to allow for age-at-birth specific control groups and time trends. For estimation, we rely on a stacked DiD estimator (Cengiz et al. 2019; Wing, Freedman, and Hollingsworth 2024) combined with a rolling window of control groups over age at birth. This estimator is flexible enough to allow not only for heterogeneous treatment effects, which eliminates the problems of forbidden comparisons and contamination, but also for control groups and time trends that are unique to each age at birth.

We apply our approach to administrative data from Germany to assess the bias in conventional event studies and to provide new insights on the heterogeneous labor market costs of motherhood. First, we document a large average impact of motherhood on earnings. For the fourth year after birth we estimate € −29,345, which is substantially larger than the corresponding result from the conventional estimation approach (€ −20,741). In relative terms, the difference amounts to 22 percentage points or 28 percent compared to the conventional estimate. Although these estimates are obtained on a specific sample of mothers – those who gave birth in West Germany between 1975 and 2021 and were in regular employment prior to childbirth – our results provide general insights beyond the specific setting. We show that conventional event studies insufficiently capture the unrealized growth in earnings that would have happened absent childbirth. Since their control group includes already-treated mothers, who are experiencing income losses or a slowdown in career progression, the counterfactual trend is biased downwards. This is the main reason why we observe an underestimation.

We further provide new evidence on how the costs of motherhood and their interpretation differ in maternal age at birth. Estimating the age-at-birth-specific effects, we find that the absolute negative impact of motherhood on earnings increases in age at first birth. We identify two major opposing factors as the sources of this pattern. On the one hand, the pre-birth earnings levels increase almost linearly until age 30 and flatten out thereafter, following the well-documented earnings path over the life cycle. This implies that leaving the labor market or reducing working hours becomes increasingly costly for older first-time mothers. On the other hand, younger first-time mothers face steeper counterfactual trends since the not-yet-mothers of close ages who make up their control groups are still at relatively early career stages and make active progression in the labor market. In relative terms, older first-time mothers lose a much smaller share of their pre-birth earnings. The negative impact of motherhood for them is composed to a large extent from losses in levels after birth rather than due to foregone growth. Younger mothers, in comparison, face the relatively larger career costs of motherhood as they miss out on the career phase with the most rapid progression. We further analyze the effects of motherhood by educational level and document that the losses are most pronounced for younger women with higher education, who stop climbing the career ladder despite having the highest potential. This pattern is consistent with the recent work by Gallen et al. (2023) who find a more pronounced negative impact of unplanned pregnancies for younger first-time mothers, pointing to a disruption in human capital accumulation at the beginning of the career as the main mechanism. The substantially different interpretations of the effects of motherhood by age underscore the importance to consider heterogeneity by age at the first birth. We point out that the negative impact of motherhood is a combination of both losses relative to the pre-birth levels and foregone earnings, with different composition across ages at birth.

Our paper contributes to several strands of literature. First, we make a methodological contribution to the vast literature on how motherhood affects women's labor market outcomes. We start by explaining how not addressing heterogeneity by age at the first birth leads to substantially biased estimates of the effects of motherhood if event studies or related methods are used (Angelov, Johansson, and Lindahl 2016; Kuziemko et al. 2018; Bütikhofer, S. Jensen, and Salvanes 2018; Fitzenberger, Sommerfeld, and Steffes 2013; Kleven, Landais, Posch, et al. 2019; Bruns 2019; Andresen and Nix 2022; Fitzenberger and Seidlitz 2023; Kleven, Landais, and Leite-Mariante 2024; Adams-Prassl, M. Jensen, and Petrongolo 2024). Building on the existing solutions to account for different treatment timing and effect heterogeneity (Cengiz et al. 2019; Callaway and Sant'Anna 2021; Sun and Abraham 2021; Wing, Freedman, and Hollingsworth 2024), we propose a new approach to estimate the labor market costs of motherhood. We extend the existing models to fit the goal of identifying motherhood effects and allow for the age-at-birth specific time trends and rolling window of control group's age at birth. Importantly, our approach yields substantially different results compared to a conventional event study, both in terms of magnitude of the effects and their development. It is flexible and can be extended to study additional dimensions of effect heterogeneity or other research questions with similar settings.

We make a further methodological contribution to the emerging body of literature on issues in DiD models and event studies with staggered adoption and heterogeneous treatment effects (de Chaisemartin and D'Haultfoeuille 2020; Goodman-Bacon 2021; Sun and Abraham 2021; Callaway and Sant'Anna 2021; Borusyak, Jaravel, and Spiess 2024; summarized by Roth et al. 2023; de Chaisemartin and D'Haultfoeuille 2023). The career costs of having children and the resulting gender inequality is a common subject of empirical work and of high relevance in public debates, but also an application where – as we demonstrate – substantial biases due to heterogeneous treatment effects are likely to be present. We follow the methods proposed by Goodman-Bacon (2021) and Sun and Abraham (2021) to illustrate in detail how the heterogeneous effects of motherhood across age at birth enter an average estimate with different weights and cause biases. Moreover, we show that the heterogeneity in our application is not just a source of biases but rather provides additional information that is economically meaningful and valuable for policymaking. These insights can be of interest beyond our specific case in settings with staggered treatment adoption and similar heterogeneity patterns, for instance health shocks or unemployment.

Furthermore, in our application of the new method, we make an empirical contribution to the large literature on the career costs of motherhood (see, among others, Adda, Dustmann, and Stevens 2017; Lundborg, Plug, and Rassmussen 2017; Blundell et al. 2021; Goldin 2021). Our results emphasize that the importance of analyzing the effects of motherhood by age at first birth goes beyond just methodological concerns. We decompose the negative effects of motherhood on earnings into realized losses in earnings levels and foregone earnings progression. We show that these two components have different importance for younger and older mothers, meaning the impact of motherhood for them is different in nature. Analyzing the effects by age at birth creates an opportunity to understand the differential responses to childbirth, which can be particularly useful when evaluating the impact of policy reforms.

The paper is organized as follows. The next section describes the data we use, then Section III gives an overview of the heterogeneity in outcomes and characteristics among mothers of different ages, which is a source of the problems in the conventional estimation approach as we discuss in Section IV. In Section V, we propose a new approach to estimate the impact of motherhood that accounts for heterogeneity by age at birth. We apply this approach in Section VI to estimate the heterogeneity-robust impact of childbirth and show how it differs in magnitudes and interpretations by age at birth. Section VII concludes.

II Data

Our analysis in this paper relies on administrative labor market data from Germany, precisely the Sample of Integrated Labor Market Biographies (SIAB, Schmucker, Seth, and vom Berge 2023; Graf et al. 2023) provided by the Institute for Employment Research. It is a two percent sample drawn from the universe of German workers who are subject to social security contributions, including administrative records of individual labor market biographies with information on employers, occupations and daily wages. We complement the SIAB with survey data from the German Socio-Economic Panel (SOEP, Goebel et al. 2019) as they provide a wider range of characteristics and draw from the full population (Section B in the Online Appendix describes the SOEP data in more detail and compares them to the SIAB).

Since the data are taken from employers' reports to the social security system they have some shortcomings. First, wages are only recorded up to the threshold for social security contributions. For wages above that ceiling we apply an imputation method that follows Dauth and Eppelsheimer (2020), who build on work by Dustmann, Ludsteck, and Schönberg (2009) and Card, Heining, and Kline (2013). Only a small share of mothers, namely 1.7 percent, have a censored wage in the pre-birth year. Second, childbirths cannot be observed directly but have to be imputed following Müller, Filser, and Frodermann (2022). This imputation utilizes the labor market absence during the maternity protection period around childbirth that mandates an employment break of at least 14 weeks. Müller, Filser, and

Frodermann show that their method identifies around 60 percent of all births in Germany. Since our population of interest is the subset of women who are in employment when giving birth, the share of identified births in our sample will be larger.[1] As suggested by Müller, Filser, and Frodermann, we restrict the sample to mothers who are at most of age 38 when giving their first childbirth to avoid false-positive identification of motherhood as the probability of labor market absence due to long-term illness increases in age.

Furthermore, the SIAB data do not include information on individuals in self-employment and civil service. As the SOEP data show, these groups do not make up a significant part of the labor market, as in the year prior to the first childbirth, only 5 percent of employed women are civil servants, and 2.4 percent are self-employed. Thus, our result should be representative of the effects of motherhood for women employed in the regular labor market.

To ensure that there is a comparable reference point for each individual, we restrict the sample to those mothers who give birth while they are employed, thus omitting births while in unemployment. Measured in the SOEP, this condition holds for more than 80 percent of mothers.[2] Further, we only keep mothers who give birth in West Germany, as data on them is consistently available for the entire sampling period.

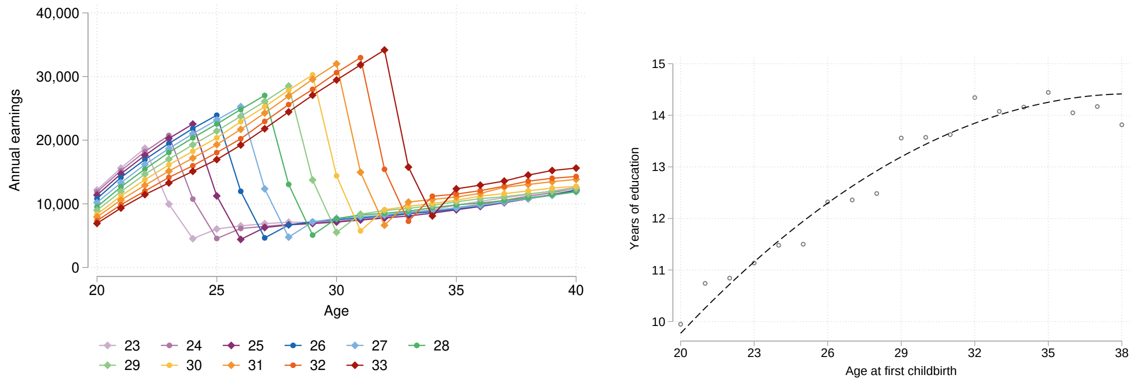
Our final sample includes around 186,000 mothers who give their first observed birth between 1975 and 2021. We aggregate the spell data to a panel by calendar years and construct annual earnings, our main outcome, as the sum of all daily wage payments from each employment spell in a year. For periods where no participation in the labor market and wage are observed, we assume earnings of zero. Throughout the paper, monetary values are given in Euro, deflated to the base year 2015. On average, mothers in the SIAB data give their first childbirth at age 28.5. In the year before, they have earned € 26,906 and have accumulated 6 years of employment experience. The majority of 68 percent owns a vocational degree, 11 percent have completed tertiary education while 15 percent have no vocational education (see Table A.1 in the Appendix).

III The Source of Problems: Heterogeneity by Age at Birth

Several papers have documented heterogeneity of maternal labor market outcomes across different ages at birth (Wilde, Batchelder, and Ellwood 2010; Adda, Dustmann, and Stevens 2017; Goldin, Kerr, and Olivetti 2022). As our paper largely builds on the observation that the characteristics and outcomes of mothers show significant differences by age at the first childbirth, this section provides a systematic overview.

[1] The applicability of the SIAB data to questions requiring the identification of mothers has further been demonstrated by various applications, for instance Schönberg and Ludsteck (2014), Collischon, Kuehnle, and Oberfichtner (2024), Huebener et al. (2024), and Boelmann, Raute, and Schönberg (2025).

[2] Boelmann, Raute, and Schönberg (2025) confirm this number. We test whether omitting women who give birth while being unemployed affects our results as part of the robustness checks in Section VI.A.



(a) Average earnings of mothers by age at first birth.

(b) Average total years of education of mothers by their age at first childbirth.

Figure 1: Heterogeneity in earnings and education among mothers by their age at the first childbirth.

Notes: Figure 1a plots the development of annual labor earnings of mothers by age at birth for the age range 20 to 40. Figure 1b shows years of education of mothers by their age at the first childbirth as a binned scatter plot with an added quadratic fit. *Source:* Own calculations based on the SIAB (Figure 1a) and SOEP (Figure 1b) data.

Figure 1a plots the development of our main outcome, annual labor earnings, over age by age at the first childbirth. It documents substantial differences in multiple aspects of mothers' earnings before as well as after birth. First, older mothers reach a higher career stage and, accordingly, higher levels of earnings before they give birth. On average, giving birth one year later relates to additional € 1,300 in annual earnings in the year prior to childbirth (see also Table A.1 in the Appendix). Second, we observe across all ages at birth that the earnings trajectories after childbirth are not constant but dynamic over time. Most apparent is the large drop in earnings right in the year of childbirth, which is substantially larger for older mothers. This is followed by continued, but smaller, losses in the first post-birth year, an uptick of earnings in the second year, and steady, but small, growth thereafter. Third, the figure shows that post-birth earnings trajectories differ substantially across age-at-birth groups. Young mothers experience little immediate recovery, while the increase in earnings between the first and second post-birth year becomes more pronounced with increasing age at birth. In addition, the growth of earnings in the longer run after birth is more pronounced for older mothers. Taken together, these patterns in labor earnings around childbirth suggest that motherhood effects are likely to be heterogeneous, i.e. different across ages at birth and dynamic over time since the event.

Among the multiple drivers of earnings, education is of particular importance. It is usually decided on already in the early stages of life and is correlated with desired fertility (Adda, Dustmann, and Stevens 2017; Doepke et al. 2023). In Figure 1b, we confirm the positive correlation between years of completed education and the age at the first childbirth. We further observe that giving birth one year later is associated with a 1.7 percentage points lower likelihood to be in the lowest education category (no vocational degree), but a 2 percentage points larger likelihood to have completed tertiary education (Table A.1 in the Appendix).

We observe similar patterns for other characteristics that describe a mother's position in the labor market. In the year prior to the first childbirth, older first-time mothers have accumulated more labor market experience (around half a year per additional year of age at the first birth) and work in firms that are larger and pay higher average wages (Table A.1 in the Appendix). They, further, earn higher hourly wages prior to the first childbirth as well as five years later, are less likely to work in part-time, and work in occupations that rank higher in terms of occupational prestige (see Figure A.1 in the Appendix).

For maternal characteristics that are related closely to fertility, the correlation with the age at the first childbirth holds as well. Older mothers spend a shorter time on parental leave (around 1.6 months per additional year of age at the first birth, see Table A.1 and Figure A.2a in the Appendix) suggesting they have retained more human capital when they re-enter the labor market. They also have fewer children in total (Figure A.2b). Both of these aspects likely play an important role in explaining the development of earnings after birth where older mothers show the steeper recovery right after birth.

We further document that older mothers have parents that are more likely to have completed high school and college education (Figure A.3 in the Appendix), which points to an inter-generational persistence of the correlation between age at the first childbirth and, on average, a more positive selection of mothers. This difference in family background also suggests that mothers of different ages at birth face different environments and barriers when it comes to receiving education and progressing in the labor market. In others words, it means that if younger mothers delay fertility by many years, it will not necessarily translate into joining the career trajectory of older first-time mothers.

Overall, age at birth incorporates information across several dimensions such as human capital investments, realized career outcomes, family-related decisions, and socio-economic background. These results provide descriptive evidence that younger and older first-time mothers differ significantly, as do their labor market outcomes and trajectories both before and after childbirth.

IV The Problem: Child Penalty Estimation under Heterogeneity

To assess the effects of the first childbirth on labor market outcomes, event studies – linear regressions that include a series of indicators for the time around childbirth, along with further covariates – are a common tool. In this section, we explain the implications of heterogeneity by age at birth for event studies and show that it causes problems if left unaccounted for. First, mechanical biases emerge as conventional event studies fail to account for the fact that mothers in each age-at-birth group are treated at different points in time and that their treatment effects are heterogeneous. Second, the heterogeneity of maternal characteristics by age at birth imposes crucial limitations on which individuals can serve as a control group.

IV.A Biases Due to Effect Heterogeneity and Staggered Adoption

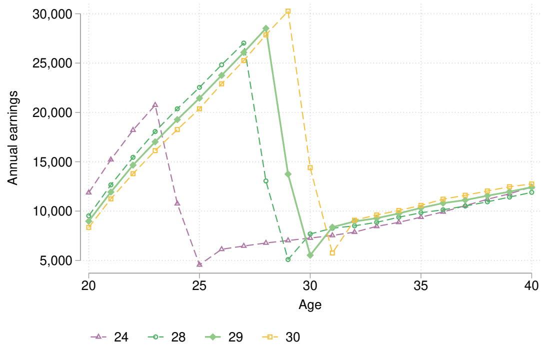
Women give birth at different ages. Thus, an event study around childbirth is an example of a research design with staggered treatment adoption. Compared to a setting where everybody receives treatment at the same time, there are two time dimensions in this staggered adoption setting – time relative to treatment and age. Importantly, these two time dimensions are aligned differently for each age-at-birth group. At each level of age, one observes mothers at different distances from their first childbirth. Some already have given birth – are already-treated – while others are still before the first birth – are not-yet treated. A binary treatment status indicator, however, pools all the age-at-birth groups together over one event time dimension. The consequence of this aggregation is that an event study estimates an average effect that is a weighted average of all possible group-by-group comparisons of a treated group with groups that are not-yet treated and groups that are already-treated (de Chaisemartin and D'Haultfoeuille 2020; Goodman-Bacon 2021; Sun and Abraham 2021; Borusyak, Jaravel, and Spiess 2024). Those comparisons to groups that already have received treatment such that their outcomes do not represent an appropriate counterfactual are typically called forbidden comparisons.

Illustration of Forbidden Comparisons To illustrate how forbidden comparisons impact the estimation of the effects of motherhood with event studies, we apply the decomposition proposed by Goodman-Bacon (2021) to a stylized example. For simplicity, we take a step back from the dynamic setting and instead consider a static one with only two periods. We focus on a single treated group, women who give birth at age 29. Our aim is to estimate the effect of childbirth on annual earnings for this group in the year of childbirth (age 29, event time 0) relative to the pre-birth year (age 28, event time −1). For illustration purposes, we include three other groups: one much earlier-treated group that gave birth at age 24, one group that gave birth in the previous year at age 28, and one not-yet-treated group that will give birth at age 30. The average earnings of these four groups over the age range 20 to 40 are plotted in Figure 2a. The figure shows that all groups exhibit similar earnings trajectories around childbirth: steady growth before birth, a sharp and large drop in period 0, smaller additional losses in period 1, some recovery in period 2, and slow growth thereafter.

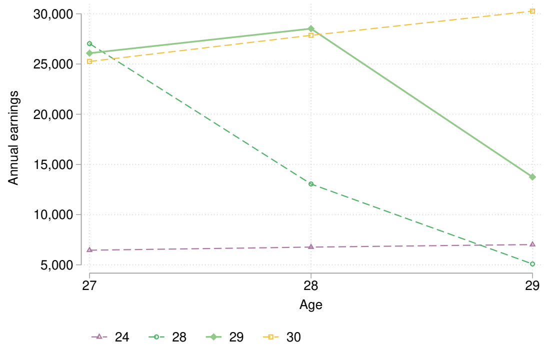
Formally, we estimate the motherhood effect in period 0 for the chosen treated group (giving birth at age 29) by fitting the following regression model to the restricted sample (age-at-birth groups 24, 28, 29, 30) and age window (ages 28 and 29):

$$Y_{ia} = \beta \times \mathbb{1}\left[a - A_i \geq 0\right] + \gamma_i + \lambda_a + \nu_{ia}. \tag{1}$$

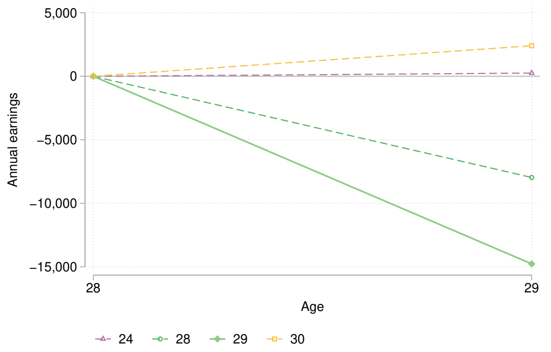
Here, Y_{ia} indicates annual earnings for mother i at age a. $\mathbb{1}\left[a - A_i \geq 0\right]$ is a treatment status indicator for mother i with age at birth A_i, which takes the value of 0 if the individual is not treated yet, switches to 1 at childbirth and stays 1 thereafter. Fixed effects for individuals (γ_i) and age (λ_a) represent the time-invariant level-differences across individuals and the general trend associated with age. Conditional on these fixed effects, the coefficient β captures the difference in the earnings trends between the treated



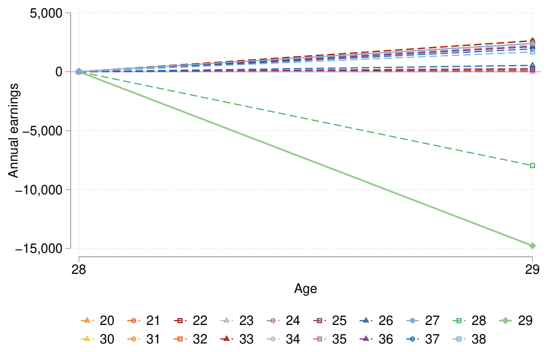
(a) Average earnings for the age range 20 to 40 (age-at-birth groups 24, 28, 29, 30).



(b) Average earnings for the age range 27 to 29 (age-at-birth groups 24, 28, 29, 30).



(c) Changes in earnings between the ages 28 to 29 (age-at-birth groups 24, 28, 29, 30).



(d) Changes in earnings between the ages 28 to 29 (age-at-birth groups 20-38).

Figure 2: Average earnings in levels and changes by age at birth over age.

Notes: The figure plots the average annual labor earnings of different groups by age at birth. The top part of the figure plots earnings in levels; its bottom part plots the change in earnings between age levels 28 and 29. Panels 2a to 2c focus on the subset of mothers who give birth at the age levels 24, 28, 29 and 30, panel 2d includes all groups from 20 to 38.
Source: Own calculations based on the SIAB (see Section II for a description).

group and the control group, as classified by the treatment status variable. In this setup, the control group's earnings trend serves as the counterfactual for how the treated group's earnings would have evolved over age in the absence of childbirth.

To better understand the dynamics of the outcome, we first look at the average annual earnings of these four age-at-birth groups over the age window of interest in Figure 2b. The only group that does not receive treatment during the observation period, women who will give birth at age 30 (dashed yellow line), continues on a smooth upward earnings trajectory over the age range from 27 to 29. The treated group – those giving birth at age 29 – is represented by a solid green line. This group shows steady earnings growth up to age 28, followed by a sharp decline in the year of childbirth (age 29). The one-year-earlier-treated group, who gave birth at age 28 (dashed green line), experiences losses in both periods, although they become smaller in magnitude in the year after birth. The much earlier-treated group, who gave birth at age 24 (dashed purple line), exhibits only minimal earnings growth between ages 27 and 29, reflecting a persistent post-birth slowdown in earnings progression.

In Figure 2c, we plot the changes in earnings for these four groups from age 28 to age 29, subtracting their levels in the pre-birth period (age 28), which in Equation 1 are captured by the individual fixed effects. This adjustment isolates the changes in the outcome from pre- to post-treatment periods, so that β in Equation 1 captures the difference in the trends of treated and control units. The comparison that researchers typically intend to make is between the treated and the non-treated groups – in this case, between the solid green and dashed yellow line – since the other two groups have already been exposed to treatment. However, this is not the only comparison being made.

Table 1: Decomposition of the average DiD estimate: Clean and forbidden comparisons.

	(1) Average	(2) Clean (to 30)	(3) All forbidden (to 24 and 28)	(4) Forbidden (to 24)	(5) Forbidden (to 28)
Treatment status	−12,731	−17,259	−10,288	−15,081	−6,800
	(116)	(135)	(128)	(136)	(156)
Age FE (Age=29)	−2,032	2,496	−4,475	318	−7,963
Person FEs	Yes	Yes	Yes	Yes	Yes
Included cohorts	24, 28, 29, 30	29, 30	24, 28, 29	24, 29	28, 29
Estimation window	28–29	28–29	28–29	28–29	28–29
Weight in the estimate from (1)		35.0 %	65.0 %	27.4 %	37.6 %
Weight in the estimate from (3)				42.1 %	57.9 %

Notes: The table reports the results from estimating the effect of childbirth on average annual labor earnings in the year of birth for the group that gives birth at age 29 following Equation (1). Column (1) reports results for the estimated average effect using a sample that includes the four groups that give birth at ages 24, 28, 29 and 30. Column (2) reports the results from clean comparisons, when only the groups 29 and 30 – treated and not-yet-treated – are included in the sample. Column (3) reports the result from forbidden comparisons, when only the groups 24, 28, and 29 – already-treated and treated – are included in the sample. Columns (4) and (5) report the coefficients separately for each of the two forbidden comparisons. Standard errors are reported in parentheses.
Source: Own calculations based on the SIAB (see Section II for a description).

Estimating Equation 1 yields an average estimate of € -12,731 (see column (1) of Table 1). However, this average estimate does not represent the difference between the trends of the treated group and the only untreated group that gives birth at 30, which equals € -17,259 (see column (2) of Table 1). The discrepancy arises, because in this staggered adoption setting the alignment of time the dimensions – time relative to treatment and age – varies across the age-at-birth groups. For those giving birth at 29, ages 28 and 29 represent one pre- and one post-birth period. At the same time, they are pre-treatment periods for the group who gives birth at 30, but post-treatment periods for groups who give birth at 24 and 28. This means that the treatment status of the three groups giving birth at 24, 28, and 30 remains unchanged throughout the estimation window. Crucially, the coefficient β captures the response in the outcome to a change in treatment status by one unit, which at the age of 29 happens only for the group that gives birth at 29.[3] The other three age-at-birth groups, for whom the treatment status does not change, then serve as control groups. Another way to see this is by noting that using individual fixed effects is equivalent to demeaning all variables by their individual pre-birth values (in this case, the value at age 28). For the already-treated groups, this results in a treatment status of 0 in both periods – just as if they were control units.

Therefore, in our example, the average estimate includes three comparisons. The clean comparison is the one where the changes in earnings of the treated group (29) and the not-yet-treated (30) one are compared. The the other two use already-treated mothers (24 and 28) as control units and, therefore, are forbidden ones, as their trends are already affected by motherhood. The group that gave birth right in the previous year (at age 28) is still experiencing substantial losses in its first post-birth year, while the group that gave birth at 24 exhibits a mostly flat trend. If the post-birth losses were constant over time and matched the losses of the group giving birth at 29, then the comparisons to already-treated mothers would yield a difference of zero and, hence, would not introduce a bias. However, since the losses differ across age-at-birth groups and decrease between the post-birth periods 0 and 1, the comparison to mothers who gave birth at 28 yields an estimate of € −6,800 (see column (5) of Table 1), which substantially understates the earnings losses for the treated group. The mostly flat earnings trend of mothers treated at age 24 leads to an underestimation as well (we estimate €−15,081, see column (4)). Using both already-treated groups as controls leads to an estimated effect of € −10,288 (column (3)), which receives a weight of almost two thirds (65 percent) in the average estimate of € −12,731 (column (1)). In contrast, the clean comparison to the not-yet-treated mothers (who give birth at 30) yields a much larger effect of € −17,259 (column 2) but receives a weight of only 35 percent in the average estimate. The resulting bias is large, € −4,528 or nearly 36 percent of the average estimate. The direction of the bias is caused by the fact that childbirth leads to earnings losses in levels and a slower earnings growth. This can be observed directly in the estimated age fixed effects, that represent the counterfactual earnings progression that would have happened from age 28 to age 29 absent children. While the clean estimate of the growth absent children is positive at € 2,496 (column (2) of Table 1), the average estimate is negative and amounts to € −2,032 (see column (1)). The difference in the age fixed effects captures the bias in the average estimate, representing the bias in the assumed counterfactual trend.

In Figure 2d, we extend the example to include all groups of mothers that give birth between the ages of 20 and 38. As before, the solid green line indicates the group treated at age 29, while the dashed green line depicts the earnings losses of the previously treated group (at age 28). A set of almost flat trends comes from the groups already treated earlier, while most of the upward sloping trends come from

[3] This is similar to the intuition that only the effects for compliers and defiers – those who change their treatment status in response to an instrument – are identified in an instrumental variable design. As Imbens and Angrist (1994, p. 470) write: "The local average treatment effect is analogous to a regression coefficient estimated in linear models with individual effects using panel data. In models with fixed effects, the data are only informative about the impact of binary regressors on individuals for whom the value of the regressor changes over the period of observation."

mothers who are not-yet treated. One additional upward trend (depicted as solid light blue line with round markers) comes from the group that was treated at age 27. In the second year after giving birth, a substantial fraction of mothers returns to the labor market, which leads to a steep uptick in the average earnings (a dynamic that is common across all groups as shown in Figure 1a). Generally, the estimate for a given age-at-birth group in a given period is composed of comparisons of the treated group's changes in earnings to mostly flat post-birth trends of the earlier already-treated groups, to a downward trend from the group treated just before, and to upward sloping trends from a recovery period and not-yet treated women. At the same time, only the group of not-yet-mothers is suitable to serve as a control one, as every other group has already received treatment. Repeating the Goodman-Bacon (2021) decomposition for this example with all age-at-birth groups shows that forbidden comparisons receive 56 percent of the weights.

Importantly, forbidden comparisons have a substantial impact on the counterfactual trend only if they receive large weights. Understanding why this is the case is insightful, as it sheds light on how widespread the problem of forbidden comparisons is when estimating the effects of motherhood. The weight that a difference-in-differences model as in Equation 1 assigns to each pairwise comparison generally depends on three components: the relative and absolute size of the compared groups as well as the timing of treatment, i.e. the position of each treated group within the time window (Goodman-Bacon 2021). The weight assigned to a comparison tends to be larger if the two compared groups are more similar in size, if the absolute size of the groups is larger, and if treatment occurs closer to the middle of the time window. The distribution of age at the first childbirth is typically bell-shaped around the average age at birth (see Figure A.6 in the Appendix for the case of Germany). With this shape, comparisons of age-at-birth groups close to the middle of the distribution, i.e. at age levels when most mothers give birth, have the largest values for all three components of the weight. Additionally, the relative size component is large for age-at-birth groups that are close to each other, since these groups are very similar in their absolute sizes. This means that comparisons to the group treated just one year earlier – who are still experiencing losses and thus provide the least suitable counterfactual scenario – tend to receive a larger weight. We confirm this in Table 1, where the comparison of the weights of the already-treated groups in columns (4) and (5) shows that mothers who gave birth one year earlier (at age 28) receive a larger weight of 37.6 percent in the average estimate (compared to 27.4 percent for those treated much earlier at age 24). Since the bell-shaped distribution of age at first childbirth is common, forbidden comparisons are prone to receive large weights in estimations of the effects of motherhood in different settings and countries. The direction of the bias depends on the outcome trajectories around childbirth. For example, steep recoveries of earnings for already-treated mothers could lead to an overestimation of the motherhood effect, while continued earnings losses or a slow-down in career progression could lead to an underestimation of the impact of motherhood.

Generalization to Conventional Event Studies and Contamination In the previous example, we simplified the analysis by considering only two time periods. Usually, researchers observe women's outcomes over a longer period of time and are interested in understanding the dynamic impact of motherhood around the event of childbirth. Allowing the effects of motherhood to vary over time formally means replacing the single treatment status indicator in Equation (1) with a set of indicators for an event time window from L_{min} to L_{max}:

$$Y_{ia} = \sum_{\substack{l=L_{min}, \\ l\neq -1}}^{L_{max}} \beta_l \times \mathbb{1}\left[a - A_i = l\right] + \gamma_i + \lambda_a + \varepsilon_{ia}. \tag{2}$$

Here, each of the dynamic treatment status indicators in $\sum_{\substack{l=L_{min} \\ l\neq -1}}^{L_{max}} \mathbb{1}\left[a - A_i = l\right]$ takes the value 1 if a mother i, who gives her first childbirth at age A_i, is l years away from A_i. For all other relative time periods $l' \in [L_{min}, l) \cup (l, L_{max}]$, it takes the value 0.

In this dynamic setup, commonly referred to as an event study, the problem of forbidden comparisons persists. The two time dimensions – time relative to treatment and age – remain aligned differently across the age-at-birth groups. At each age, there are mothers who have already given birth ($l > 0$), mothers who give birth at this exact age ($l = 0$), and those who have not yet become mothers ($l < 0$). At each age, an event time indicator takes the value of 1 for mothers who are exactly l years away from their first birth. For those mothers who are not l, but $l' \neq l$ years away from their first birth, the respective indicator takes the value of 0. Importantly, the indicator equals 0 not only for women who have not yet become mothers ($l' \in [L_{min}, -1]$), but also for those who have already given birth ($l' \in [1, L_{max}]$), resulting in forbidden comparisons.

In addition to forbidden comparisons, a second issue can introduce a bias into event studies estimates – contamination (Sun and Abraham 2021). When replacing the binary treatment indicator with a set of indicators for each relative time period around treatment, the effect for a given period is estimated conditional on the effects from all other periods. In other words, the estimated average effect of being in a specific period relative to childbirth is subtracted from the outcomes of each age-at-birth group at that period. If the treatment effects of motherhood are homogeneous – that is, constant across relative time periods and age-at-birth groups – then this subtraction will correctly account for the effect of being treated a number of periods away. However, with heterogeneous treatment effects, a single average is not sufficient to account for the heterogeneous effects across the age-at-birth groups. The result of this conditional estimation of heterogeneous effects over multiple relative time periods is that the differences in effects from the other periods l' enter the estimate of the effect in l and contaminate it. In Figure A.4 in Appendix A, we use the decomposition proposed by Sun and Abraham (2021) to illustrate contamination when estimating the effects of motherhood. It confirms that the estimates of the treatment effects in a given relative time period contain the effects from other periods, leading to a contamination of the estimates.

The results discussed above have important implications for the commonly applied event studies around childbirth that follow the model proposed by Kleven, Landais, and Søgaard (2019):

$$Y_{iat} = \sum_{\substack{l=L_{min},\\ l\neq -1}}^{L_{max}} \beta_l \times \mathbb{1}[a - A_i = l] + \lambda_a + \delta_t + u_{iat}. \tag{3}$$

This specification includes fixed effects for age and calendar year (λ_a and δ_t). Often, researchers include additional individual fixed effects to account for the time-invariant differences across individuals as in Equations 1 and 2[4]. As we have demonstrated above, this conventional event study setup is susceptible to both forbidden comparisons and contamination. Therefore, its results are prone to being neither a numerically correct representation of the actual average effect of motherhood nor an interpretable weighted average of the effects of motherhood across multiple age-at-birth groups.

Biases in Re-scaled Estimates Additionally, many applications follow the approach by Kleven, Landais, and Søgaard (2019) to re-scale the event study estimates to relative terms. As a re-scaling benchmark, they use a counterfactual outcome that is obtained by predicting the outcomes using only the fixed effects from Equation (3). The goal of this re-scaling is to capture the ratio between the effect of childbirth and the predicted outcome without children, allowing to interpret the absolute losses in terms of counterfactual earnings. However, the counterfactual outcome itself is prone to be biased, as it is derived from an estimation that includes already-treated units in the control group.

Table 1 provides an illustration of this issue. The coefficient of the age fixed effect represents the estimated earnings progression for the control group. The clean estimate of the age fixed effect in column (2) amounts to € 2,496, whereas the average estimate in column (1) is negative at € −2,032, even though both aim to capture the progression of annual earnings for women without children. The reason for the diverging age fixed effects between the clean and the average estimate is that the latter additionally includes the downward and flat trends of already-treated mothers. Therefore, this conventional re-scaling approach is prone to including biases in both the numerator and the denominator. We further analyze this issue in the Online Appendix, Section D, where we show how the composition of the denominator changes over relative event time.

IV.B Control Group Choice Under Heterogeneity

The problems discussed above can be considered mechanical ones, in the sense that already-treated individuals enter the control group while, in fact, it should include only non-treated ones. However, at the core of an event study research design lies the choice of a valid control group composed of non-treated individuals. Event studies recover unbiased treatment effects under two assumptions which are typically

[4] Notably, in any event study model that includes treatment status indicators, these indicators automatically classify individuals into treated and control groups. Therefore, including unit fixed effects is necessary to account for the pre-treatment level differences between the groups, which would otherwise be mistakenly attributed to the treatment effect.

referred to as no anticipation and parallel trends (Sun and Abraham 2021). Assuming no anticipation means that no one should be affected by treatment before the event of treatment. Assuming parallel trends means that the trend of the control group should represent the counterfactual development that the treated group would have had absent treatment. Especially the second assumption is crucial for the choice of a control group as it requires general comparability of outcome trajectories between the treatment and the control group in the untreated state. Therefore, the pronounced differences in maternal outcomes by age at birth have important implications for choosing a control group, as they threaten the comparability of mothers across age-at-birth groups. In the following, we discuss the validity of the most common and available control groups with respect to the heterogeneity in maternal characteristics by age at birth.

Men If men are used as control units, event studies capture the effect of having a child on the outcomes of mothers under the assumptions that men are unaffected by childbirth and that, absent children, women would have the same outcome trajectories as men. In practice, these assumptions are unlikely to hold. First, there are gender differences in career paths because of various kinds of discrimination and different experiences and decisions that can have an effect already prior to birth, for instance when choosing a college major or an occupation (Blau and Kahn 2017). Moreover, after childbirth, being a father can have both positive and negative effects on outcomes of men. Negative ones can arise in societies that are more gender-equal where the burden of raising children is shared more equally between both parents. Positive effects for fathers are possible as well, for example, if specialization within couples re-distributes incentives to further invest in the labor market career to husbands. Goldin, Kerr, and Olivetti (2022), for instance, conclude the existence of a "fatherhood premium" that is larger for more time-intensive occupations which points towards an increase in productivity due to a focus on market work (similar to the findings on male marital wage premiums that is documented, among others, by Antonovics and Town 2004).

In Figure A.5 in Appendix A, we plot the average earnings trajectories of fathers by age-at-parenthood groups and document no visible impact of childbirth in the German setting. However, we observe that there is a positive correlation between the levels and the growth of labor market outcomes of fathers and age at birth. This suggests that heterogeneity by age at parenthood should be taken into account as well when estimating the impact of fatherhood using event studies. Consequently, this also holds if the gap in the effect of parenthood is of interest or if the analysis makes comparisons within couples (as for instance in Angelov, Johansson, and Lindahl 2016; Andresen and Nix 2022).

Childless Women Using childless women as control units requires to assume comparability and common trends for mothers and non-mothers, as well as no anticipation of not receiving treatment. Both assumptions are unlikely to hold given the endogeneity of fertility decisions. Lundborg, Plug, and Rassmussen (2017) utilize the special case of in vitro fertilization success to estimate the career effects of childbirth conditional on receiving IVF treatment, and Gallen et al. (2023) use failing oral contraception to identify the effects of unplanned pregnancy and childbirth on labor market outcomes. However, in a general sample of mothers, it is challenging to distinguish a random component from selection into fertility. Employing childless women as control group does not account for the voluntary decision of being childless for career reasons, while Steinhauer (2018) documents interrelations between gender roles, childlessness, and choices in the labor market. Moreover, it neglects the heterogeneity among the childless individuals, some of whom decide to not have children, whereas others tried to but could not have them. Outcome trajectories of childless women, who might know they will remain childless (for instance, due to medical reasons) or explicitly plan to do so, are very unlikely to represent the counterfactual trajectories of mothers in the absence of children.

Not-Yet-Mothers Restricting the control group to not-yet-treated women allows to account for the selection of women into having children instead of remaining childless. This group is used in the dynamic matching approaches by Fitzenberger, Sommerfeld, and Steffes (2013) and Fitzenberger and Seidlitz (2023). Yet, all not-yet-treated women include much older first-time mothers. As we show in Section III, the timing of childbirth is correlated with some important characteristics of mothers and their labor market outcomes, both before and after birth. On average, older first-time mothers have a higher socio-economic status, reflected in higher earnings, educational levels, occupational ranks, and other labor market outcomes. Therefore, the outcomes of much older first-time mothers cannot serve as a counterfactual for younger mothers, such that, in their entirety, not-yet-treated units are unsuitable as a control group.

However, the group of not-yet-treated women also includes those mothers whose age at first birth is close to that of each treated group. In Section III, we show that age at childbirth and characteristics relevant for labor market outcomes are strongly correlated with and, importantly, change smoothly in age at birth. Therefore, bringing together women of close ages at birth in the treatment and control groups can establish a basis for the comparability of their outcome trajectories. This idea is key to the new solution that we introduce in the next section.

V The Solution: Heterogeneity-Robust DiD with Rolling-Window Control Groups

We build on the preceding descriptive evidence on the systematic heterogeneity by age at birth and the analysis of the problems it causes to inform our choice of a new approach to model, identify, and estimate the effects of motherhood. In the following, we first extend the potential outcomes model for a DiD design under staggered treatment adoption to allow heterogeneity in both treatment effects and counterfactual time trends by age at birth. Second, we specify unique control groups for each age-at-birth group that consist of only the closest ages at first birth and formulate identifying assumptions for the target parameters, the average effect of becoming a mother specific for mothers of each age-at-birth group. Lastly, we build an estimation procedure based on a stacked DiD estimator (Cengiz et al. 2019; Wing, Freedman, and Hollingsworth 2024), extending it to allow for heterogeneous control groups and time trends by age at birth.

Potential Outcomes Model We start with a potential outcomes model for a DiD model with staggered treatment adoption. We build on models that allow for heterogeneous treatment effects under staggered adoption of treatment, proposed by de Chaisemartin and D'Haultfoeuille (2020), Sun and Abraham (2021), Callaway and Sant'Anna (2021), and Wing, Freedman, and Hollingsworth (2024), and adapt them to suit the goal of identifying the effects of motherhood on women's outcomes.

We observe the outcomes Y_{ia} of women indexed as $i \in \{1, ..., N\}$ over age $a \in \{a_{min}, ..., a_{max}\}$. Age at the first birth, which we denote as $A_i \in \{A_{i,min}, ..., A_{i,max}\}$, serves as treatment timing dimension. Childbirth is a binary and absorbing event, and eventually all women become mothers in our setting. Treatment timing is thus the only source of variation in treatment status. Women who become mothers at the same age $A_i = A$ make up the groups in terms of treatment timing.

Accordingly, we define $Y_{ia}(0)$ and $Y_{ia}(1)$ as the potential outcomes of a woman i at age a in the untreated and treated states. The realized outcomes are connected to the potential outcomes through the treatment timing:

$$Y_{ia} = Y_{ia}(0) \times \mathbb{1}[a < A_i] + Y_{ia}(1) \times \mathbb{1}[a \geq A_i].$$

We model the potential outcomes in the untreated and treated state as

$$\begin{aligned} Y_{ia}(0) &= \gamma_i + \theta_{a,A} + \varepsilon_{ia}, \\ Y_{ia}(1) &= Y_{ia}(0) + \beta_{a-A,A} + \nu_{ia}. \end{aligned}$$

As in common DiD research designs, we model time-invariant unobserved heterogeneity, γ_i, on the individual level. Following Sun and Abraham (2021), Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfoeuille (2020), and Wing, Freedman, and Hollingsworth (2024), we allow the treatment effects, $\beta_{a-A,A}$, to vary with age at birth A and over time since birth $a - A$, i.e. to be both heterogeneous and dynamic. Diverging from the mentioned models, we do not assume a common time trend, but rather allow the age trend, $\theta_{a,A}$, to be heterogeneous by age at birth A. This modeling choice is crucial to estimate the effects of motherhood since mothers who give birth at different ages have different characteristics (see Section III) and, thus, would likely have had different trajectories absent children. ε_{ia} and ν_{ia} refer to random error terms.

The parameter that we are interested in identifying with this model is the causal effect of motherhood on women's outcomes, specific to age at birth and period relative to birth. Formally, we specify the target parameter as the Average Treatment Effect on the Treated (ATT) for a given treatment timing group (A) and period after birth ($a - A$), similar to the parameters in Sun and Abraham (2021) and Callaway and Sant'Anna (2021):

$$ATT_{a-A,A} = \mathbb{E}[Y_{ia}(1) - Y_{ia}(0)|A_i = A].$$

Identifying Assumptions In DiD models with staggered adoption, identification of the parameter $ATT_{a-A,A}$ builds on two key assumptions: parallel trends absent treatment specific to treatment timing and restricted anticipation prior to treatment (Callaway and Sant'Anna 2021; Sun and Abraham 2021).

In our setting, the first assumption means that, absent childbirth, the outcomes of the treated and control groups, specific to age at birth, would have evolved in parallel. This assumption requires choosing a suitable control group for each age-at-birth group, such that their specific outcome trajectory (captured by the age trend $\theta_{a,A}$) indeed represents the counterfactual development for treated mothers who give birth at age A. To make this choice, we leverage the strong correlation between age at birth and the relevant outcomes and decisions of mothers, as well as their smooth change in age at birth, documented in Section III. This correlation suggests that, conditional on having close ages at birth, outcome trajectories are comparable. Therefore, we specify the age-at-birth-specific control groups as the pre-birth observations of those mothers who give birth at slightly older ages. For each group that gives birth at age $A_i = A$, we include just a few next groups of mothers before they give birth, G, in each age-at-birth-specific control group such that it consists only of mothers with close ages at birth $A_i \in \{A + 1, A + G\}$. Formally, this translates into assuming that the average change in outcomes of the treated group with $A_i = A$ in the post-birth periods ($a \geq A_i$) and of their respective control group with $A_i \in \{A + 1, A + G\}$ in the pre-birth periods ($a < A_i$) would have been the same in the absence of children:

$$\mathbb{E}[Y_{ia}(0) - Y_{i,A-1}(0)|A_i = A, a \geq A_i] = \mathbb{E}[Y_{ia}(0) - Y_{i,A-1}(0)|A + 1 \leq A_i \leq A + G, a < A_i].$$

Intuitively, we assume that women, who give birth at a certain age, absent children would have followed the outcome trajectory of the women who give birth at slightly older ages. Compared to the approaches proposed by Sun and Abraham (2021), Callaway and Sant'Anna (2021), and Wing, Freedman, and Hollingsworth (2024), we allow the control groups and counterfactual trajectories to differ by age at birth and assume parallel trends only group-specifically and within a narrow window of age at birth. This allows us to account for the fact that mothers with different ages at birth give birth at different life and career stages.

Second, we assume that there is only limited anticipation of treatment in the pre-birth periods $a < A_i$:

$$\mathbb{E}\left[Y_{ia}(0)|A_i = A\right] = \mathbb{E}\left[Y_{ia}(0)|A + 1 \leq A_i \leq A + G\right].$$

Specifically, we rule out dynamic anticipation within a range of close ages at birth. This is a weaker assumption than the standard version of no anticipation, which is too strong to be plausible in the context of motherhood. In general, we can expect two types of potential anticipation in this setting – long-term and short-term one. Long-term anticipation can arise from plans for future fertility that influence women's decisions earlier in life – for instance human capital investments or occupation choices – well in advance of childbirth (Adda, Dustmann, and Stevens 2017). Short-term anticipation can emerge when women make fertility decisions based on their current career outcomes or adjust labor market choices in response to pregnancy. Our choice to include only women with close ages at birth in the control group and to assume no anticipation within this range of ages at birth mitigates long-term anticipation, as we group together only women who are similar in terms of fertility timing and the associated early-life decisions. As for short-term anticipation, Callaway and Sant'Anna (2021) discuss the possibility to adjust this assumption if researchers want to allow for anticipation within a fixed number of periods before childbirth. For example, this adjustment can be useful if there is the suspicion that women tend to give birth following a promotion or a slow-down in their career progression. In this case, with p pre-treatment periods where anticipation is suspected, no anticipation should be assumed for the periods $\{a < A_i - p\}$ instead of $\{a < A_i\}$.

Estimation Procedure To estimate the $ATT_{a-A,A}$ under the identifying assumptions stated above, we need to employ an estimation procedure which allows not only for heterogeneous effects but also for heterogeneous counterfactual trends by age at birth. While de Chaisemartin and D'Haultfoeuille (2020), Sun and Abraham (2021), and Callaway and Sant'Anna (2021) propose estimators that allow for effect heterogeneity, they use never-treated, all not-yet-treated, or the last-treated groups as control groups. This means that they assume a common counterfactual trend for all groups treated at different points in time, which is borrowed from one control group. However, we model the counterfactual trend and the control group to be different for each age-at-birth group, reflecting that these groups are different and would have had different trajectories absent children. Therefore, we employ an estimation procedure which gives us the required flexibility – namely, stacked DiD estimator, which was introduced by Cengiz et al. (2019) and further developed by Wing, Freedman, and Hollingsworth (2024). We extend the standard stacked DiD estimator by using a rolling window of control groups over age at birth, which allows us to specify unique control groups of the closest ages at birth for each age-at-birth group of mothers.

In the stacked DiD estimator, each group of women becoming mothers at a certain age is treated as a separate sub-event. Similar to Sun and Abraham (2021) and Callaway and Sant'Anna (2021), it estimates group-specific effects that can be aggregated to obtain an average estimate. Allowing the effects to vary for each age-at-birth group eliminates forbidden comparisons and contamination. It solves the misalignment of event time and age across age-at-birth groups by aligning – stacking – them by event time. Furthermore, the stacked DiD design allows a flexible definition of control groups and, consequently, heterogeneous time trends by age at birth, which is crucial for our approach. For each age-at-birth group, we define a specific control group in which we include only the pre-birth observations of not-yet-treated mothers who are the closest in terms of age at the first childbirth.

Formally, each group of women becoming mothers belongs to a separate sub-event S_i and construct according sub-datasets that only include the treated group – those who gives birth at age $A_i = A$ – and their assigned control observations – the pre-birth observations of those who give birth at later ages $A_i \in \{A+1, A+G\}$. Appending the sub-datasets vertically by event time creates a stacked dataset over all sub-events, in which all age-at-birth groups are aligned by event time. We then fit a saturated linear regression, where a standard two-way fixed effects (TWFE) regression is interacted with sub-event indicators, to the stacked dataset:

$$Y_{ia} = \sum_{A=A_{min}}^{A_{max}} \sum_{\substack{l=L_{min},\\ l\neq -1}}^{L_{max}} \Big(\beta_{A,l} \times \mathbb{1}[A_i = A] \times \mathbb{1}[a - A_i = l] + \gamma_{i,A} + \theta_{a,A} + \varepsilon_{ia,A}\Big) \times \mathbb{1}[S_i = A]. \tag{4}$$

In the equation, Y_{ia} indicates the outcome of mother i at age a. Accordingly, the model includes a set of event-time indicators $\mathbb{1}[a - A_i = l]$ that identify when a mother with $A_i = A$ is l years away from her first childbirth at age A_i. The indicator function $\mathbb{1}[A_i = A]$ identifies the treated units within a sub-event, i.e. it takes the value of 1 if unit i belongs to the treated group in sub-event and takes a value of 0 if unit i belongs to the control group in sub-event. The indicator function $\mathbb{1}[S_i = A]$ identifies each treated group of mothers (with $A_i = A$) along with the pre-birth observations of the assigned control units (with $A_i \in \{A+1, A+G\}$ and $a < A_i$), allowing the coefficient $\beta_{A,l}$ to vary by sub-event. The fixed effects for individual and age, $\gamma_{i,A}$ and $\theta_{a,A}$, are allowed to vary by sub-event as well. This fully interacted regression estimates the coefficients as if we ran separate TWFE regressions for each group of women becoming mothers at a certain age using the stacked dataset. We cluster the standard errors at the individual level to account for correlation of the error term over time.

Even though the age-at-birth-specific estimates discussed so far provide additional information, it is often of interest to calculate an average effect across all groups. To this end, we follow Wing, Freedman, and Hollingsworth (2024) and Sun and Abraham (2021) and weight the age-at-birth-specific estimates by the sample shares of each group:

$$\hat{\beta}_l = \sum_{A=A_{min}}^{A_{max}} \frac{N_A}{N} \times \hat{\beta}_{A,l}, \tag{5}$$

where N_A indicates the number of observations per group and N the total number of observations. The result is a weighted average estimate over all age-at-birth groups. Accordingly, the standard error of the average can be obtained as a linear combination of the age-at-birth-specific standard errors using the same weights.

To be able to recover a correctly weighted average ATT_l, where $\hat{\beta}_l$ in each time period is informed by the full set of age-at-birth-specific estimates, it is crucial to ensure that each group in the stacked dataset is observed over all pre- and post-birth periods (Wing, Freedman, and Hollingsworth 2024), i.e. that the panel is balanced. This additional restriction has implications for the estimation window. The time horizon for which our approach allows to estimate the effects of motherhood is defined by the ranges of age at the first birth and observed age available in the data as well as the number of age-at-birth groups included in the control group (G). The number of not-yet-treated age-at-birth groups which the researcher is willing to include in the control group (G) defines the upper limit of the time horizon the estimation can cover ($L_{max} = G - 1$), since for $l > G - 1$ there are no not-yet-treated observations left to serve as control units. The requirement to construct a balanced panel implies that G also defines the oldest age-at-birth group, for which it is possible to estimate the effects, ($A_{max} = A_{i,max} - G$). Depending on how many pre-birth periods (L_{min}) the researcher wants to include to test the parallel trends assumption, the youngest group in a balanced panel is defined by the first observed age level ($A_{min} = a_{min} - L_{min}$).

Effects in Relative Terms A common last step when estimating the effects of motherhood is to re-scale the estimates to relative terms. The outcome Y is usually in absolute terms which allows to keep values of zero during times of non-participation in the sample. At the same time, researchers often report the effects of motherhood in percentage terms to give a measure of the effects' magnitude and to improve the comparability across different settings, for instance across countries. To allow comparisons across studies and settings, the re-scaling benchmark should be unbiased, consistent, and transparent. We propose to re-scale the age-at-birth-specific estimates with the age-at-birth-specific pre-birth levels. In contrast to the conventional re-scaling method that we discuss in Section IV.A, this approach correctly accounts for the different relative magnitudes of losses that depend on the initial levels, remains constant over time, and has the clear interpretation of indicating losses with respect to the pre-birth levels.

Flexibility The estimation approach we present here does not require additional information compared to a conventional event study. Thus, it can be used with the data from existing applications, for instance, to make comparisons of the labor market costs of motherhood across countries. Moreover, the stacked DiD design is flexible and can be adjusted to study further heterogeneity in the effects of motherhood. Effect heterogeneity for certain groups of mothers – for instance, by pre-birth occupation or skill-level – can be investigated by interacting the model with indicators for the respective group; in Section VI.B we provide an example that differentiates by education. The effects of having consecutive children and their timing, highlighted as important factor by Adams-Prassl, M. Jensen, and Petrongolo (2024), can be studied if the second birth is set as an event of interest. Dynamics by calendar time can be explored using multi-way stacking, i.e. by additionally considering sub-events by the calendar time dimension for each age-at-birth group. Our approach can also be used to study other research questions in staggered adoption settings with similar heterogeneity patterns by age at the event: for instance, the impact of migration, health shocks, unemployment, marriage, or divorce.

VI New Insights: Heterogeneous Effects of Motherhood

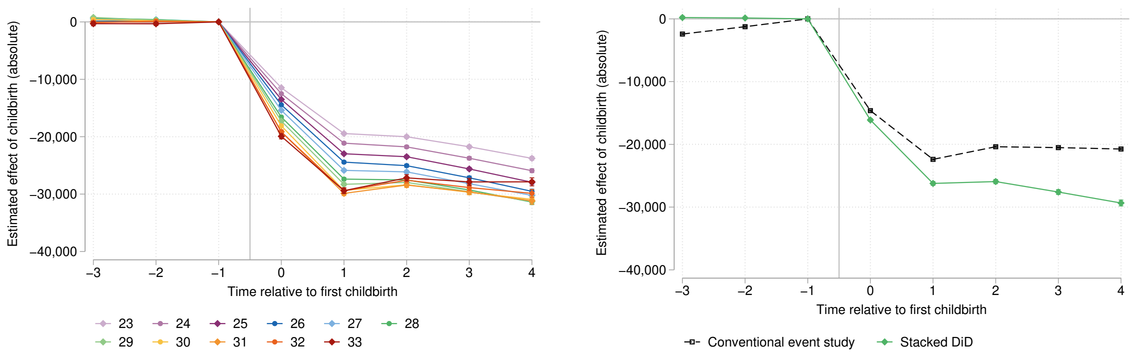
Our approach to estimate the effects of motherhood by age at birth presented in Section V not only helps to correct potential biases but also allows to study effect heterogeneity. In this section, we illustrate this key advantage of our stacked DiD design by applying it to administrative German labor market data (see Section II for details about the SIAB data). Our findings confirm substantial heterogeneity in the effects of motherhood across ages at birth, both in levels and trends. Importantly, we show that this heterogeneity has a meaningful and policy-relevant interpretation – one that is lost if only an average estimate is shown. Furthermore, we highlight the importance to account for heterogeneity by age at birth when estimating the average motherhood effect, as our approach yields estimates that differ substantially from those obtained using a conventional event study.

In our application, we include five not-yet-treated age-at-birth groups in each control group ($G = 5$) to exclude much older, and therefore less comparable, first-time mothers from the control groups (we also provide a test for the sensitivity of the results to the choice of G). This allows us to estimate effects for up to four years after childbirth ($L_{max} = 4$). With three included pre-birth periods ($L_{min} = -3$) and data on the age-at-birth groups 20 to 38 ($A_i \in \{20, ..., 38\}$), this means we estimate the effects of motherhood for women who give the first birth between ages 23 and 33 ($A_{min} = 23$ and $A_{max} = 33$).

VI.A Effects of Motherhood by Age at Birth and on Average

Figure 3a presents our main results on the impact of motherhood on annual labor earnings by age at first childbirth. We document substantial variation in the effects across age-at-birth groups, with differences that emerge already in the year of childbirth and widen over time after childbirth. By the fourth year after birth, the estimated earnings losses range from € 23,769 for the youngest mothers in the sample to € 31,389 for those who gave birth at age 28 (see Table A.2 in Appendix A for all numerical results and standard errors). Earnings losses increase with age at birth up to age 28, remain around their level at this age up to age 31, after which they decline slightly for the oldest groups, € 29,870 at age 32 and € 27,860 at age 33. These findings confirm the substantial heterogeneity in the effects of motherhood by age at birth and highlight the importance of accounting for it. Besides the large differences in the post-birth periods, we document flat pre-trends for all age-at-birth groups. This provides supporting evidence for the key identifying assumption of parallel trends, as defined in Section V, that the earnings growth is similar for each age-at-birth group of mothers and the associated control groups absent treatment. We also find no evidence of short-term anticipation in the pre-birth periods, which is in line with the descriptive analysis of the average age-at-birth-specific outcomes before birth in Figure 1a.

Aggregating the age-at-birth-specific results to an average estimate (following Equation 5 and using the sample shares of each age-at-birth group, documented in Figure A.6, as weights) allows a comparison with the results from a conventional event study (see Equation 3). Figure 3b presents the estimated motherhood effects from both approaches. Compared to the conventional method, our stacked DiD approach yields effects that are larger in absolute terms starting from the year of childbirth, with the



(a) Age-at-birth-specific estimates.

(b) Weighted average estimate compared a to conventional event study estimate.

Figure 3: Motherhood effects in annual earnings after the first childbirth (along with 95 percent confidence intervals).

Notes: The figure plots the estimated effects of motherhood (along with 95 percent confidence intervals) on annual labor earnings of women after the first childbirth following Equation (4). Figure 3a reports the age-at-birth-specific estimates, Figure 3b reports the weighted average of the age-at-birth-specific estimates following Equation (5) and compares that to the results of a conventional event study as in Equation (3). Table A.2 in Appendix A reports the corresponding coefficient estimates.
Source: Own calculations based on the SIAB (see Section II for a description).

gap between the estimates widening over time. By the fourth year after childbirth, the estimate from the stacked DiD approach indicates an average negative impact of around € 29,345, which is € 8,604 larger than the corresponding estimate from the conventional event study. Relative to the unbiased estimate, this difference amounts to 29 percent.

In this comparison, the different dynamics of the estimates are particularly important. Aggregating the age-at-birth-specific estimates from our approach results in flat average pre-trends and a downward sloping post-birth average trajectory. In contrast, the conventional event study estimates exhibit a clear upward-sloping pre-trend followed by a flat post-birth trend. These patterns are in line with the expectations from our analysis in Section IV.A, which concludes that conventional event studies are prone to biases since they include already-treated mothers in the control group. Since mothers typically experience either earnings losses or a slowdown in earnings growth following childbirth, this translates into a downward bias in the counterfactual earnings growth and generates both an upward-sloping pre-trend and a flat post-birth trend. Our approach avoids these forbidden comparisons by constructing clean control groups composed of women who have not yet given birth and continue to experience earnings growth in the absence of children. Combined with the slowdown in earnings from having a child that mothers experience, this results in a downward-sloping post-birth trend in the average estimates. As the control group earnings continue to accumulate over time, the difference between the two approaches widens accordingly, driven by the conventional method's underestimation of the counterfactual earnings growth.

Robustness Checks We perform several checks to assess the robustness of our findings. Table C.1 in the Online Appendix lists the results. First, we test whether our results are driven by mothers with particularly high earnings. To this end, we either drop observations with earnings above the 99th percentile (part II in Table C.1) or all those mothers whose earnings are censored at the social security threshold in the year prior to childbirth (part III in Table C.1) from the sample. Both checks lead to average coefficient estimates that are numerically in absolute terms a bit smaller (in year four after birth, € −27,606 and € −27,539, i.e. a reduction of around 6 percent compared to the main specification), but exhibit a similar pattern as the main specification. These smaller estimates are expected, since we primarily exclude mothers who lose substantially more when reducing employment or leaving the labor market.

Next, we assess the importance of unemployment for our results. Our estimation sample conditions on being employed in the year prior to the first childbirth to ensure that we observe the impact of childbirth on labor market careers for a consistent set of mothers. As this approach allows for being unemployed after, but not at, birth, one might be concerned that post-birth unemployment biases the treatment groups' outcomes downwards while unemployment is less likely to affect the pre-birth observations in the control groups. In part IV of Table C.1, we assess this issue by dropping all mothers who have at least one post-birth year with more than 30 days receipt of unemployment benefits. The results remain almost unchanged with an absolute earnings loss of € −29,263 in the fourth year after birth. Part V of the table takes a different approach to assess the effect of unemployment on our estimates. Rather than excluding unemployed mothers, it explicitly includes also mothers who are unemployed when they give birth. As the estimation sample is now more heterogeneous and includes a number of mothers with no or rather low earnings, we observe in both the conventional estimation and the stacked DiD approach an upward shift of the results to levels that are smaller in absolute terms (in year four after birth, € −19,501 for the conventional estimation and € −27,150 in the stacked DiD). Importantly, however, our previous conclusions continue to hold. We still observe a substantial difference between the estimation approaches (which amounts to € 7,649 or 28 percent of the stacked DiD estimate), as well as a flat post-birth trend for the conventional estimation, whereas the stacked DiD approach continues to estimate a downward sloping earnings trajectory after childbirth.

In addition, we explore the impact of including mothers who give birth at a closer or further away age in the age-at-birth-specific control groups. Our baseline specification uses mothers who give birth up to five years later than each respective age-at-birth group as control units, which allows to estimate effects for up to four post-birth periods. For instance, for the age-at-birth group 26, the control group consists of mothers giving birth at ages 27 to 31. We now repeat our estimation multiple times, each time using not multiple, but just one age-at-birth group of not-yet-mothers as control units; specifically we iterate from control mothers who give birth three up to eight years later, i.e. we estimate effects for two up to seven post-birth periods. This test shows that the average estimates over all age-at-birth groups (plotted in Figure C.1a in the Online Appendix) change only little when using control groups that are either closer or more further away in terms of age at birth. Going beyond average estimates, we also show age-at-birth-specific estimates for ages 26, 29, and 32. Here, we find that varying the control group leads to negligibly different results for younger age-at-birth groups (Figure C.1b). For older mothers, however, using not-yet mothers whose age at birth is further away as control units increases the variation in the estimates (Figures C.1c and C.1d). Compared to estimates that use very close not-yet mothers, those obtained with mothers who are much older at birth yields coefficient estimates that are smaller in absolute terms. This highlights that with higher age the control group is more likely to be affected by the common flattening in the age-earnings profile over the life cycle. The practical implication of this is that if one aims to estimate the effect of childbirth for mothers who are thus far mostly unaffected by a flattening of their earnings, but uses mothers who already experience it as control units the counterfactual earnings trend is more prone to be underestimated, leading to smaller point estimates. This observation is likely to hold across different datasets and settings. However, as the development of earnings over the life cycle can differ across settings, the choice of the oldest age at birth group in the control group ($A + G$) remains specific to the respective application. A formal approach to choosing G requires structurally modeling the selection into fertility timing (e.g. as in Adda, Dustmann, and Stevens 2017), which is outside the scope of this paper and most applied work on motherhood effects.

Lastly, using the SOEP data, we assess if the absence of civil servants and self-employed in the SIAB could lead to a bias in our estimates. That would occur if a substantial number of women move from regular employment to civil service or self-employment after childbirth, such that they show up with zero earnings in the SIAB while they are, in fact, working. We observe, that the share of civil servants remains almost unchanged at around 5 percent from the year prior to childbirth to the fourth post-birth year. Over the same period, the share of self-employed women almost doubles from 2.4 to 4.6 percent. However, of those self-employed after birth only 26 percent (i.e. 0.57 percentage points of the increase in the share of self-employed) were employees in the pre-birth year, which makes it unlikely that moves to self-employment have more than a marginal impact on the estimates.

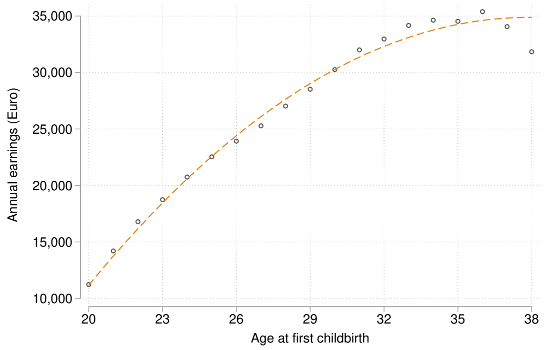
VI.B Heterogeneous Effects by Age at Birth

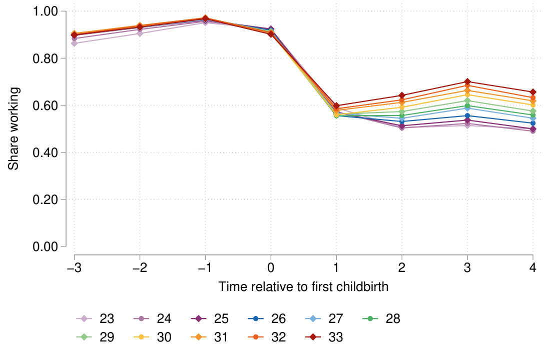
Beyond addressing methodological issues, the heterogeneous effects documented above reflect systematic differences across women and in their reaction to childbirth. Thus, they offer an opportunity to gain a better understanding of the impacts of motherhood.

Sources of Differential Effects of Motherhood Ex ante, the pattern shown in Figure 3a is ambiguous. Older first-time mothers, who are at a later career stage, may have stronger incentives to return to work and can be expected to earn more if they do. At the same time, they mechanically lose more than younger mothers if they reduce their working hours or (temporarily) leave the labor force. In the figure, we observe that earnings losses in levels increase almost linearly for mothers who give the first birth in their twenties, which reverses at age 31, with absolute losses slightly decreasing for older mothers. Moreover, examining the trajectories over time, we observe downward-sloping earnings trends for younger mothers that flatten with higher age at birth. To better understand these patterns, we decompose the age-at-birth-specific estimates into losses in levels and losses in growth and show their implications for the interpretation of the effects.

Starting with losses in levels, Figure 4a plots pre-birth annual labor earnings by age at first childbirth. It shows steep increases over the twenties, followed by a flattening during the thirties. This pattern reflects the well-documented life-cycle earnings path, with a first decade that is marked by rapid progression and diminishing returns to tenure afterwards (see, among others, Bagger et al. 2014). The most substantial earnings progression typically happens during the twenties, corresponding to the first post-education career stage. As a result, older first-time mothers – having already completed this high-growth career stage – have more to lose when they leave the labor force or reduce working hours. The resulting incentive to return to work earlier for older mothers is reflected by the average employment rates around childbirth that are plotted in Figure 4b. While older first-time mothers are more likely to return to the labor market, the differences in post-birth participation across ages at birth are relatively small (around 10 percentage points with a mean of slightly below 60 percent). Thus, the share of older mothers who leave the labor market is still substantial, such that we observe levels of earnings losses that increase in age at the first birth.

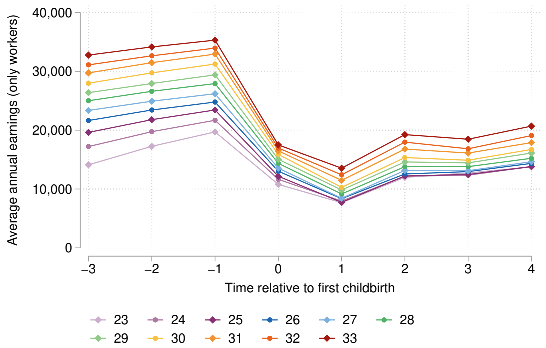
As a second step, we study the heterogeneity of losses in earnings growth by age at birth. In contrast to the losses in levels, foregone earnings progression stems from the different outcome trajectories of mothers and control women and accumulates over time. In Figure 4c, we start by plotting the earnings of working mothers across ages at birth, isolating the intensive margin. Although we find that older first-time mothers have higher levels of earnings if they return to work, there are no substantial differences in the realized post-birth growth rates across ages at birth. Therefore, the differential dynamics of earnings losses that we document in Figure 3a can be caused by the differences in participation rates documented above as well as by different counterfactual earnings growth. To inspect this, in Figure 4d we compare the earnings trajectories of the age-at-birth-specific treatment and the control groups. The control groups' trajectories depict the development of earnings that would have happened absent childbirth for each age-at-birth group. For treated mothers after birth, we observe overall little earnings growth that is slightly larger for older mothers, whereas counterfactual earnings growth is clearly steeper for younger mothers and flattens out for older ones. This matches the generally steeper earnings earnings path at younger ages discussed above. Contrary to a conventional estimation, our rolling window of control groups over age at birth is able to capture this. For younger first-time mothers, the control groups consist of younger not-yet-mothers, who are in the earlier career phase with more rapid progression. The older control individuals for the older first-time mothers are at a later career stage, where we observe a slow-down in progression. In contrast to our findings for levels, where older mothers lose more, foregone progression plays a larger role for younger first-time mothers, for whom childbirth disrupts climbing a concave career ladder at its beginning. This further highlights that it is important to

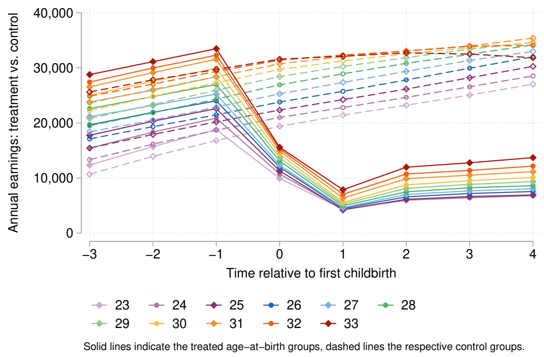




(a) Average annual labor earnings in the year prior to the first childbirth.

(b) Average employment rates.





(c) Average annual labor earnings conditional on working.

(d) Earnings trajectories of the treated age-at-birth groups and their respective control groups.

Figure 4: Decomposition of heterogeneity in absolute motherhood effects by age at birth.

Notes: The figure decomposes the heterogeneity of effects of childbirth on annual earnings by age at birth. Figure 4a plots the pre-birth levels of annual earnings. Figures 4b and 4c plot how labor market participation and earnings for participating mothers develop around the first birth. Figure 4d plots the development of earnings for each treated age-at-birth group (depicted as solid lines) and their associated control groups (dashed lines).
Source: Own calculations based on the SIAB (see Section II for a description).

construct the control groups by age at the first birth to account for the different career stages when children are born and the resulting heterogeneous losses in earnings growth.

Both patterns – larger losses in levels for older first-time mothers and larger forgone progression for younger ones – can only be observed if the analysis is conducted by age at childbirth. Using the same control group for different ages at birth is not able to account for these patterns and providing solely an average estimate hides these important differences. Crucially, the different interpretations of the negative effects of childbirth by age have meaningful implications for policy analysis. Losses in levels for older mothers point to difficulties with keeping up and advancing a previously built career, while younger mothers miss out on the opportunity to accumulate human capital and build a career in the first place. Therefore, reducing the labor market costs of motherhood for mothers of different ages may require different policies, which has to be taken into account when designing effective institutions and interventions.

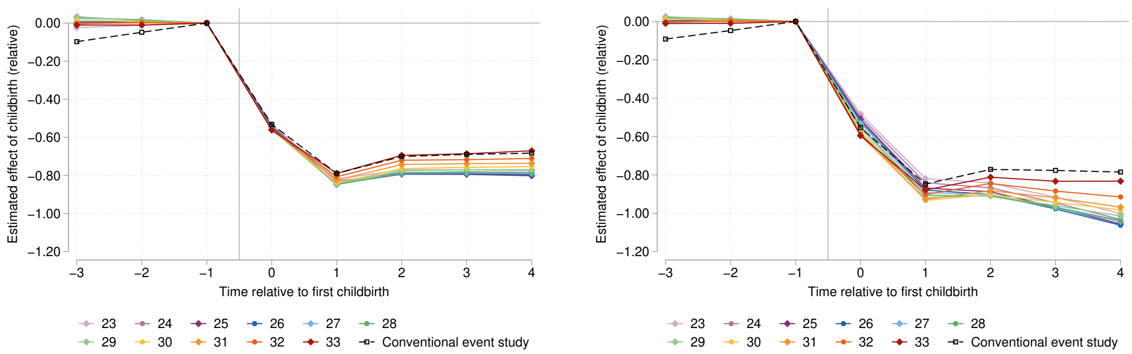
Effects of Motherhood by Age at Birth and Education The role of differential counterfactual earnings trajectories becomes even more apparent when we estimate the effects of motherhood separately by three educational levels – no vocational training, vocational training, or university education. The results are presented on the left-hand side of Figure A.7 in the Appendix. The right-hand side of this figure plots the age-at-birth- and education-specific age fixed effects, i.e. a representation of foregone earnings progression that is taken directly from the stacked DiD estimation.

In line with expectations, estimated earnings losses increase in absolute terms for better educated mothers. In addition, we also observe that the degree of heterogeneity by age at birth increases substantially in education, such that the effects for mothers with no higher education are more homogeneous across ages at birth than for mothers with university-level education. Along with the estimated effects, the variation in the age fixed effects increases as well. At all education levels, steeper age fixed effects indicate that foregone earnings progression plays a larger role for younger than for older mothers. This is by far most pronounced for highly educated women. Those women with high levels of human capital have the largest progression to forego when they leave the labor market or move to part-time jobs. Among them, the younger ones show the steepest upward-sloping counterfactual earnings trends and accordingly the steepest downward-sloping effects of childbirth, highlighting that especially highly-educated women who become mothers at younger ages miss out on the phase of the most rapid career progression for which they had the highest potential.

VI.C Effects of Motherhood in Relative Terms

Understanding the two components of the negative impact of motherhood – losses in levels and foregone growth – is particularly important when re-scaling the estimates to relative terms. This is a common step in the literature as it helps to compare the impact of motherhood across countries or policy settings. The conventional approach, that is usually applied to the estimates from Equation (3), is to divide the coefficients on the relative event-time dummies by a counterfactual outcome, i.e. the predicted value from the regression based only on the included fixed effects. As discussed in Section IV.A, effect heterogeneity introduces biases not only to the estimated treatment effects, but also to the estimate of the counterfactual outcome. However, even if effect heterogeneity is correctly accounted for, the counterfactual outcome, by construction, is not constant over event time since it includes both the static pre-birth outcome level and foregone progression as a dynamic component. Consequently, dividing the treatment coefficients by this counterfactual removes the dynamic component, which potentially leads to a misinterpretation of the relative magnitude and the dynamics of the effects of motherhood. We demonstrate this in Figure 5a, which shows the results of applying the conventional re-scaling method to our estimates from Figure 3a. As expected, this approach removes much of the observed heterogeneity across the age-at-birth groups, as it eliminates both differences in pre-birth levels and in counterfactual growth. The result of removing the dynamic component – foregone progression – is that the re-scaled estimates display a flat trend over time.

As an alternative, we propose in Section V to re-scale the estimates using the pre-birth outcome levels specific to each age-at-birth group. This allows to account for the level differences across age-at-birth groups, recognizing that a fixed earnings loss has a relatively greater impact for younger mothers with lower pre-birth earnings. It also reflects the common earnings path over the life cycle, is easy to report and, as a pre-birth measure, remains constant over time, allowing for a consistent and transparent interpretation. Figure 5b presents the corresponding results where we divide the estimates from Figure 3a by the age-at-birth-specific pre-birth earnings levels. Consistent with our previous analysis, the heterogeneity in relative losses now follows a distinctly different pattern compared to the absolute losses. Relative to their pre-birth earnings, older first-time mothers experience substantially smaller losses than younger mothers. This pattern is only partly driven by differential losses in the levels of earnings, but to a large extent stems from the different dynamics in the progression of earnings. The latter drives a fast divergence of the relative losses across the age-at-birth groups, where the trajectories for mothers below the age of 30 show steeper downward slopes and, in year four after childbirth, even exceed a value of one. This result underlines our earlier finding that, for younger mothers, the foregone growth component is particularly pronounced in the negative impact of motherhood and accumulates over time.



(a) Re-scaled with counterfactual outcomes.

(b) Re-scaled with pre-birth levels.

Figure 5: Effects of motherhood on annual earnings in relative terms.

Notes: The figure plots the estimated losses in annual earnings after the first childbirth (following Equation (4) and reported in Figure 3a and Table A.2), re-scaled to percentages. Figure 5a re-scales by dividing with the counterfactual outcomes, Figure 5b re-scales by dividing with the age-at-birth-specific earnings levels in the pre-birth year. Both figures contrast the age-at-birth-specific results with results from a conventional event study (following Equation (3)) that are re-scaled according to the respective method in each figure. Table A.3 in Appendix A reports the corresponding numerical values.
Source: Own calculations based on the SIAB (see Section II for a description).

Our estimation approach yields results that are significantly different from the conventional event study for both re-scaling methods. The re-scaled effects estimated with a conventional event study either lie on the upper bound (for the conventional re-scaling) or even outside (when re-scaling with the pre-birth levels) the range of our age-at-birth-specific estimates. As we predicted in Section IV and have documented for the effects in absolute terms, our approach corrects the downward bias and yields substantially larger estimates. Comparing the impacts of motherhood in year four after childbirth, the conventional event study re-scaled with the pre-birth levels yields an average loss of almost −79 percent, whereas our stacked DiD approach re-scaled similarly yields estimates between −83 and −106 percent with an average over all age-at-birth groups of around −100 percent. The resulting average difference between the approaches is 22 percentage points or 28 percent of the conventional estimate. Given that the effects of family-related policies are often found to be small in magnitude (Olivetti and Petrongolo 2017), a discrepancy of this size can substantially change the results of policy evaluations that rely on event studies.

VII Conclusion

Estimating motherhood effects with event studies has quickly become a widely used and valuable tool to assess the labor market effects of childbirth on mothers. In this paper, we show that the considerable amount of heterogeneity, both in maternal outcomes and characteristics, by age at first childbirth leads to these event studies producing substantially biased estimates. Applying the insights from the recent literature on difference-in-difference models, we show that event studies with childbirth as treatment are likely to make forbidden comparisons, i.e. to compare just-treated mothers to mothers who already had their child. Moreover, they can suffer from contamination, i.e. the estimates of the effect of motherhood in one relative time period can contain the motherhood effect from other time periods. Since heterogeneity by age is common among mothers across different countries and settings, such biases are likely widespread across many existing applications.

Instead of relying on conventional event studies, we propose a novel approach to estimate the labor market impact of motherhood. We extend the potential outcomes model to allow for both heterogeneous effects and time trends by age at birth. We use a stacked DiD design to estimate the impact of motherhood specifically for each age at birth. We also construct the control groups for each age at birth group, using observations only from those not-yet mothers who are the closest ones in age. This rolling window of control groups ensures that treatment and control groups are comparable, which is a crucial argument to justify the plausibility of the parallel trends assumption. The flexibility of our estimation approach allows to adapt it to other research questions with similar settings, for instance the impact of migration or health shocks at different stages of the life cycle.

Our application revisits the estimation of motherhood effects in the German labor market. It shows that a conventional event study substantially underestimates the negative impact of motherhood on earnings after the first childbirth. Furthermore, we demonstrate that recognizing the heterogeneity by the age at which mothers give birth is not only important to eliminate biases, but also allows to gain a more complete understanding of the career costs associated with motherhood. Younger first-time mothers experience larger relative labor market costs of children, that primarily stem from unrealized progression in the crucial early career stage, whereas losses in levels play the larger role for older mothers. Being at a different stage of life and career and selecting into giving birth earlier or later are likely correlated with differential responses to policies. Our results, therefore, highlight that it is important to analyze the effects of motherhood and related policies by age at childbirth.

References

Adams-Prassl, Abi, Mathias Jensen, and Barbara Petrongolo (2024). "Birth Timing and Spacing: Implications for Parental Leave Dynamics and Child Penalties". IZA Discussion Paper No. 17438 (cit. on pp. 5, 23).

Adda, Jérôme, Christian Dustmann, and Katrien Stevens (2017). "The career costs of children". In: *Journal of Political Economy* 125.2, pp. 293–337 (cit. on pp. 6–8, 21, 27).

Andresen, Martin Eckhoff and Emily Nix (2022). "What Causes the Child Penalty? Evidence from Adopting and Same-Sex Couples". In: *Journal of Labor Economics* 40.4, pp. 971–1004 (cit. on pp. 5, 17).

Andrew, Alison, Oriana Bandiera, Monica Costa-Dias, and Camille Landais (2021). "Women and men at work". In: *IFS Deaton Review of Inequalities* (cit. on p. 2).

Angelov, Nikolay, Per Johansson, and Erica Lindahl (2016). "Parenthood and the Gender Gap in Pay". In: *Journal of Labor Economics* 34.3, pp. 545–579 (cit. on pp. 5, 17).

Antonovics, Kate and Robert Town (2004). "Are All the Good Men Married? Uncovering the Sources of the Marital Wage Premium". In: *American Economic Review* 94.2, pp. 317–321 (cit. on p. 17).

Bagger, Jesper, François Fontaine, Fabien Postel-Vinay, and Jean-Marc Robin (2014). "Tenure, Experience, Human Capital, and Wages: A Tractable Equilibrium Search Model of Wage Dynamics". In: *American Economic Review* 104.6, pp. 1551–1596 (cit. on p. 28).

Blau, Francine D. and Lawrence M. Kahn (2017). "The Gender Wage Gap: Extent, Trends, and Explanations". In: *Journal of Economic Literature* 55.3, pp. 789–865 (cit. on pp. 2, 17).

Blundell, Richard, Monica Costa-Dias, David Goll, and Costas Meghir (2021). "Wages, Experience, and Training of Women over the Life Cycle". In: *Journal of Labor Economics* 39.S1, S275–S315 (cit. on p. 6).

Boelmann, Barbara, Anna Raute, and Uta Schönberg (2025). "Wind of Change? Cultural Determinants of Maternal Labor Supply". In: *American Economic Journal: Applied Economics* 17.2, pp. 41–74 (cit. on p. 7).

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (2024). "Revisiting Event Study Designs: Robust and Efficient Estimation". In: *The Review of Economic Studies* 91.6, pp. 3253–3285 (cit. on pp. 3, 5, 10).

Bruns, Benjamin (2019). "Changes in Workplace Heterogeneity and How They Widen the Gender Wage Gap". In: *American Economic Journal: Applied Economics* 11.2, pp. 74–113 (cit. on p. 5).

Bundesagentur für Arbeit, ed. (2021). *Klassifikation der Berufe 2010 – überarbeitete Fassung 2020* (cit. on p. 37).

Bütikhofer, Aline, Sissel Jensen, and Kjell Salvanes (2018). "The role of parenthood on the gender gap among top earners". In: *European Economic Review* 109, pp. 103–123 (cit. on p. 5).

Callaway, Brantly and Pedro H. C. Sant'Anna (2021). "Difference-in-Differences with multiple time periods". In: *Journal of Econometrics* 225, pp. 200–230 (cit. on pp. 3–5, 19–21).

Card, David, Jörg Heining, and Patrick Kline (2013). "Workplace Heterogeneity and the Rise of West German Wage Inequality". In: *The Quarterly Journal of Economics* 128.3, pp. 967–1015 (cit. on p. 6).

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer (2019). "The Effect of Minimum Wages on Low-Wage Jobs". In: *The Quarterly Journal of Economics* 134.3, pp. 1405–1454 (cit. on pp. 3–5, 19, 21).

Collischon, Matthias, Daniel Kuehnle, and Michael Oberfichtner (2024). "Who Benefits from Cash-for-Care? Effects of a Home Care Subsidy on Maternal Employment, Childcare Choices, and Children's Development". In: *The Journal of Human Resources* 59.4, pp. 1011–1051 (cit. on p. 7).

Dauth, Wolfgang and Johann Eppelsheimer (2020). "Preparing the sample of integrated labour market biographies (SIAB) for scientific analysis: a guide". In: *Journal for Labour Market Research* 54 (cit. on p. 6).

de Chaisemartin, Clément and Xavier D'Haultfoeuille (2020). "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects". In: *American Economic Review* 110.9, pp. 2964–2996 (cit. on pp. 3, 5, 10, 19, 21).

– (2023). "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey". In: *The Econometrics Journal* 26 (3), pp. C1–C30 (cit. on pp. 3, 5).

Doepke, Matthias, Anne Hannusch, Fabian Kindermann, and Michèle Tertilt (2023). "The economics of fertility: A new era". In: *Handbook of the Economics of the Family*. Vol. 1. 1. Elsevier, pp. 151–254 (cit. on p. 8).

Dustmann, Christian, Johannes Ludsteck, and Uta Schönberg (2009). "Revisiting the German Wage Structure". In: *The Quarterly Journal of Economics* 124.2, pp. 843–881 (cit. on p. 6).

Fitzenberger, Bernd, Aderonke Osikominu, and Robert Völter (2005). "Imputation Rules to Improve the Education Variable in the IAB Employment Subsample". In: *IAB FDZ Methodenreport* 3 (cit. on pp. 46, 47).

Fitzenberger, Bernd and Arnim Seidlitz (2023). "Changing Fertility and Heterogeneous Motherhood Effects: Revisiting the Effects of a Parental Benefits Reform" (cit. on pp. 5, 18).

Fitzenberger, Bernd, Katrin Sommerfeld, and Susanne Steffes (2013). "Causal effects on employment after first birth – A dynamic treatment approach". In: *Labor Economics* 25, pp. 49–62 (cit. on pp. 5, 18).

Gallen, Yana, Juanna Schrøter Joensen, Eva Rye Johansen, and Gregory F. Veramendi (2023). "The Labor Market Returns to Delaying Pregnancy". SSRN 4554407, `http://dx.doi.org/10.2139/ssrn.4554407` (cit. on pp. 5, 17).

Goebel, Jan, Markus M. Grabka, Stefan Liebig, Martin Kroh, David Richter, Carsten Schröder, and Jürgen Schupp (2019). "The German Socio-Economic Panel (SOEP)". In: *Jahrbücher für Nationalökonomie und Statistik* 239.2, pp. 345–360 (cit. on pp. 6, 46).

Goldin, Claudia (2021). *Career & Family: Women's Century-Long Journey Toward Equity*. Princeton University Press (cit. on p. 6).

Goldin, Claudia, Sari Pekkala Kerr, and Claudia Olivetti (2022). "When the Kids Grow Up: Women's Employment and Earnings across the Family Cycle". NBER Working Paper No. 30323 (cit. on pp. 7, 17).

Goodman-Bacon, Andrew (2021). "Difference-in-differences with variation in treatment timing". In: *Journal of Econometrics* 225.2, pp. 254–277 (cit. on pp. 3, 5, 10, 14).

Graf, Tobias, Stephan Grießemer, Markus Köhler, Claudia Lehnert, Andreas Moczall, Martina Oertel, Alexandra Schmucker, Andreas Schneider, Stefan Seth, Ulrich Thomsen, and Philipp vom Berge (2023). *Weakly anonymous Version of the Sample of Integrated Labour Market Biographies (SIAB) – Version 7521 v1*. Ed. by Research Data Centre of the Federal Employment Agency (BA) at the Institute of Employment Research (IAB). DOI: `http://dx.doi.org/10.5164/IAB.SIAB7521.de.en.v1`. The data access was provided via on-site use at the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) and subsequently remote data access. (cit. on p. 6).

Huebener, Mathias, Jonas Jessen, Daniel Kuehnle, and Michael Oberfichtner (2024). "Parental Leave, Worker Substitutability and Firms' Employment". In: *The Economic Journal* 135, pp. 1467–1495 (cit. on p. 7).

Imbens, Guido W. and Joshua D. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects". In: *Econometrica* 62.2, pp. 467–475 (cit. on p. 13).

Kleven, Henrik, Camille Landais, and Gabriel Leite-Mariante (2024). "The Child Penalty Atlas". In: *Review of Economic Studies*. Forthcoming (cit. on p. 5).

Kleven, Henrik, Camille Landais, Johanna Posch, Andreas Steinhauer, and Josef Zweimüller (2019). "Child Penalties across Countries: Evidence and Explanations". In: *American Economic Review: Papers & Proceedings* 109, pp. 122–126 (cit. on p. 5).

Kleven, Henrik, Camille Landais, and Jakob Egholt Søgaard (2019). "Children and Gender Inequality: Evidence from Denmark". In: *American Economic Journal: Applied Economics* 11.4, pp. 181–209 (cit. on pp. 2, 16, 50).

Kuziemko, Ilyana, Jessica Pan, Jenny Shen, and Ebonya Washington (2018). "The Mommy Effect: Do Women Anticipate the Employment Effects of Motherhood?" NBER Working Paper No. 24740 (cit. on p. 5).

Lundborg, Petter, Erik Plug, and Astrid Würtz Rassmussen (2017). "Can Women Have Children and a Career? IV Evidence from IVF Treatments". In: *The American Economic Review* 107.6, pp. 1611–1637 (cit. on pp. 6, 17).

Müller, Dana, Andreas Filser, and Corinna Frodermann (2022). "Update: Identifying mothers in administrative data". In: *IAB FDZ Methodenreport* 01/2022 (cit. on pp. 6, 7).

Olivetti, Claudia and Barbara Petrongolo (2017). "The Economic Consequences of Family Policies: Lessons from a Century of Legislation in High-Income Countries". In: *Journal of Economic Perspectives* 31.1, pp. 205–230 (cit. on p. 32).

Petrongolo, Barbara and Maddalena Ronchi (2020). "Gender gaps and the structure of local labor markets". In: *Labour Economics* 64, p. 101819 (cit. on p. 2).

Roth, Jonathan, Pedro H.C. Sant'Anna, Alyssa Bilinski, and John Poe (2023). "What's trending in difference-in-differences? A synthesis of the recent econometrics literature". In: *Journal of Econometrics* 235.2, pp. 2218–2244 (cit. on pp. 3, 5).

Schmucker, Alexandra, Stefan Seth, and Philipp vom Berge (2023). "Sample of Integrated Labour Market Biographies (SIAB) 1975–2021". In: *IAB FDZ Datenreport* 02/2023 (cit. on p. 6).

Schönberg, Uta and Johannes Ludsteck (2014). "Expansions in Maternity Leave Coverage and Mother's Labor Market Outcomes after Childbirth". In: *Journal of Labor Economics* 32.3, pp. 469–505 (cit. on p. 7).

Steinhauer, Andreas (2018). "Working Moms, Childlessness, and Female Identity". SciencesPo, LIEPP Working Paper No. 79 (cit. on p. 18).

Sun, Liyang and Sarah Abraham (2021). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". In: *Journal of Econometrics* 225.2, pp. 175–199 (cit. on pp. 3–5, 10, 15, 17, 19–22, 41).

Wilde, Elizabeth Ty, Lily Batchelder, and David T. Ellwood (2010). "The Mommy Track Divides: The Impact of Childbearing on Wages of Women of Differing Skill Levels". NBER Working Paper No. 16582 (cit. on p. 7).

Wing, Coady, Seth M. Freedman, and Alex Hollingsworth (2024). "Stacked Difference-in-Differences". NBER Working Paper No. 32054 (cit. on pp. 4, 5, 19–22).

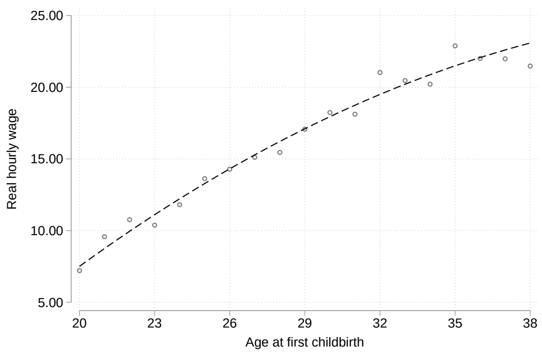
Appendix

A Additional Tables and Figures

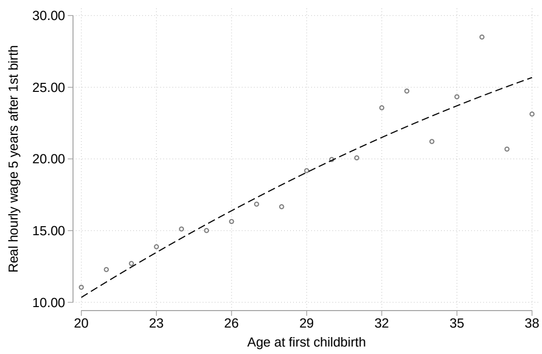
Table A.1: Correlation of mothers' characteristics with their age at the first childbirth along with summary statistics.

	Coefficient	Standard Error	Mean	SD
Pre-birth outcomes				
Annual labor earnings	1,340	9	26,906	16,595
Education				
No vocational degree	−0.017	0.000	0.154	0.361
Vocational degree	−0.001	0.000	0.682	0.466
University degree	0.020	0.000	0.112	0.316
Years in employment	0.552	0.002	6.010	3.978
Occupational rank	1.774	0.016	80.180	28.979
Complex tasks	0.019	0.000	0.161	0.368
Firm size	28	2	802	3,508
Firm-level mean daily wage	2.677	0.021	98.245	39.405
Post-birth outcomes				
Months on parental leave	−1.569	0.029	38.183	55.177

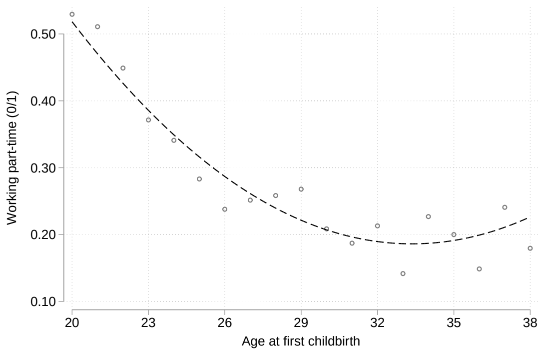
Notes: The table lists results from linear regressions of mothers' characteristics on their age at the first childbirth along with standard errors as well as means and standard deviations of the respective outcome. $N = 186{,}229$; the average age at the first childbirth is 28.53 (SD = 4.47). All characteristics except for parental leave taking are measured in the year prior to the first childbirth. Occupational rank is measured as the median wage by occupation at the 4-digit level (defined in the KldB 2010 by Bundesagentur für Arbeit 2021), including observations from both women and men, irrespective of parenthood. Task complexity is measured as an indicator that takes the value 1 for occupations that require complex or highly complex tasks as recorded in the fifth digit of the KldB 2010. Firm-level mean wages and firm size are provided in the SIAB data.
Source: Own calculations based on the SIAB (see Section II for a description).



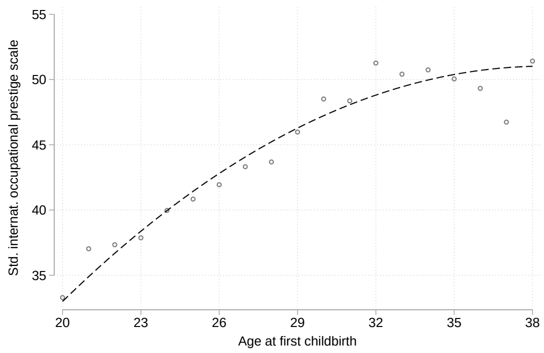
(a) Real hourly wage in the pre-birth year.



(b) Real hourly wage five years after the first birth.



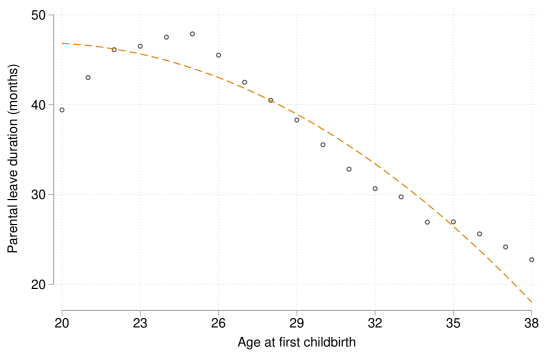
(c) Employment in part-time five years after the first birth.



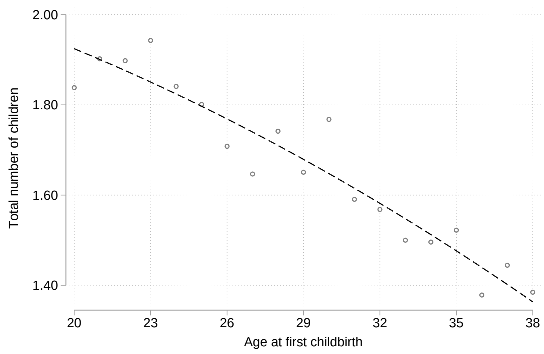
(d) Standard International Occupational Prestige Scale (range from 6 to 78) in the pre-birth year.

Figure A.1: Labor market outcomes of mothers by age at first childbirth.

Notes: The figure plots several labor market outcomes of mothers by their age at the first childbirth as binned scatter plots with an added quadratic fit.
Source: Own calculations based on the SOEP (see Online Appendix B for a description).



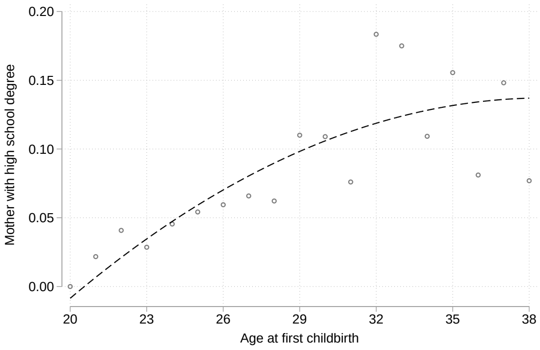
(a) Employment break duration after the first childbirth (SIAB).



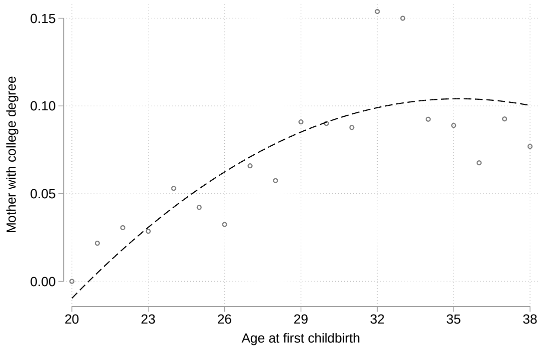
(b) Total number of children (SOEP).

Figure A.2: Employment break duration and total number of children by age at the first childbirth.

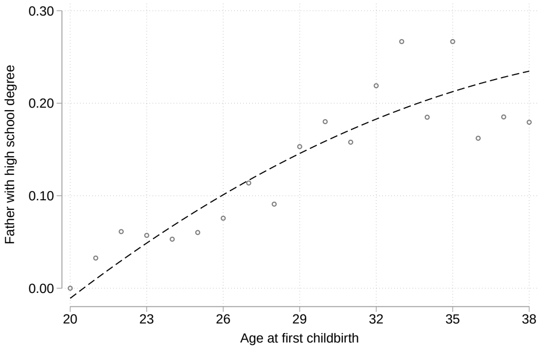
Notes: The figure plots the average duration in months of the employment break after the first childbirth and the total number of children by mothers' age at the first childbirth as binned scatter plots with an added quadratic fit.
Source: Own calculations based on the SIAB and SOEP data (see Section II for a description).



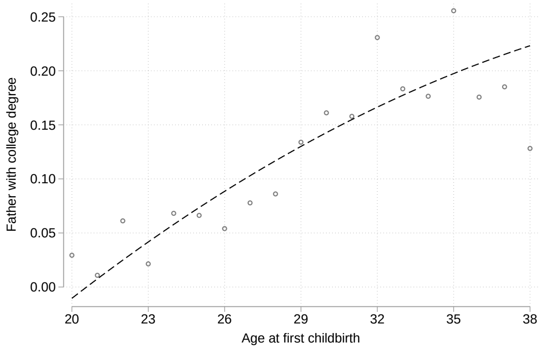
(a) Probability of mother having a high school degree.



(b) Probability of mother having a college degree.



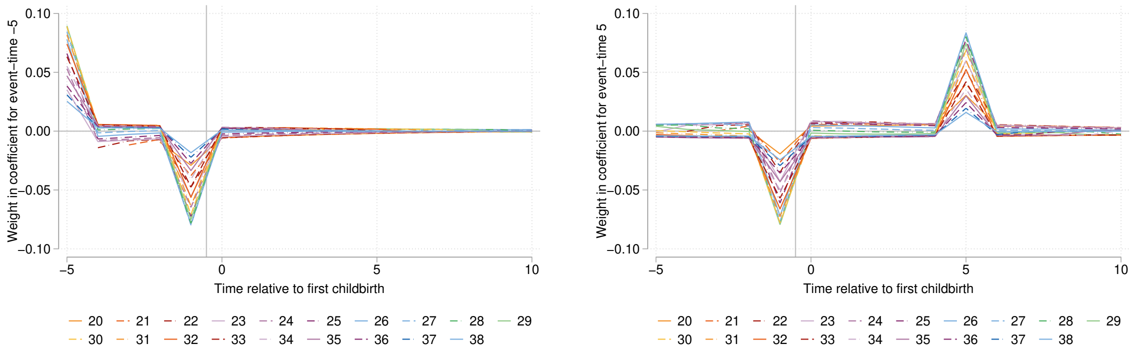
(c) Probability of father having a high school degree.



(d) Probability of father having a college degree.

Figure A.3: Educational attainment of mothers' parents by age at first childbirth.

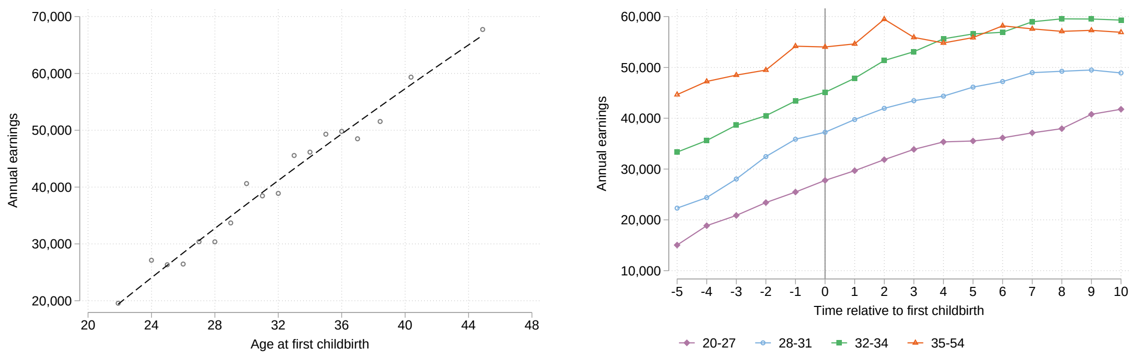
Notes: The figure plots the probability of mothers' parents to own a high school (German "Abitur") and tertiary-level education degree by their age at the first childbirth as binned scatter plots with an added quadratic fit.
Source: Own calculations based on the SOEP (see Online Appendix B for a description).



(a) Weights for the period −5 coefficient. (b) Weights for the period +5 coefficient.

Figure A.4: Sun and Abraham (2021) decomposition of weights: Contamination from other periods.

Notes: The figure plots the weights each other relative time period (on the x-axis) receives in a conventional child penalty estimation (as in Equation (3)) of the effects for the relative time periods −5 (left-hand panel) and +5 (right-hand panel). The weights are calculated with the `eventstudyweights` Stata module provided by Sun and Abraham (2021).
Source: Own calculations based on the SIAB (see Section II for a description).



(a) Real annual earnings in the pre-birth year. (b) Real annual earnings around childbirth.

Figure A.5: Average earnings of fathers by age at parenthood.

Notes: The left figure plots annual earnings of fathers in the year prior to their first childbirth against their age at the first childbirth as a binned scatter plot with an added quadratic fit. The right figure plots average annual earnings of fathers by years relative to their first birth for the four quartiles of the distribution of age at first birth. The first quartile includes fathers aged 20-27 at first birth, the second those aged 28-31, the third those aged 32–34 and the fourth those from 35-54. All monetary values are in Euro, deflated to the base year 2015.
Source: Own calculations based on the SOEP data (see Online Appendix B for a description).

Table A.2: Average and age-at-birth-specific effects on annual earnings of mothers.

	Time relative to 1st birth						
	−3	−2	0	1	2	3	4
Stacked DiD Estimation							
Average	213	141	−16,114	−26,227	−25,931	−27,599	−29,345
	(36)	(27)	(41)	(54)	(62)	(73)	(86)
Age-at-birth group							
23	−305	−183	−11,468	−19,450	−20,003	−21,752	−23,769
	(95)	(73)	(103)	(126)	(141)	(151)	(170)
24	−31	57	−12,508	−21,125	−21,771	−23,749	−25,920
	(89)	(66)	(99)	(120)	(133)	(146)	(171)
25	16	57	−13,509	−22,968	−23,486	−25,616	−27,935
	(84)	(63)	(96)	(116)	(130)	(146)	(173)
26	131	134	−14,454	−24,427	−25,043	−27,156	−29,517
	(83)	(63)	(96)	(119)	(135)	(152)	(189)
27	543	392	−15,400	−25,861	−26,128	−28,180	−30,226
	(90)	(68)	(103)	(126)	(144)	(163)	(201)
28	659	401	−16,576	−27,388	−27,533	−29,277	−31,389
	(101)	(75)	(112)	(137)	(155)	(177)	(223)
29	780	293	−17,271	−28,293	−27,965	−29,585	−31,249
	(114)	(84)	(124)	(152)	(171)	(194)	(254)
30	408	206	−18,109	−29,352	−28,372	−29,686	−30,915
	(127)	(96)	(138)	(168)	(190)	(216)	(258)
31	−38	34	−19,130	−29,875	−28,420	−29,610	−31,173
	(144)	(115)	(163)	(195)	(215)	(241)	(307)
32	−103	134	−19,177	−29,440	−27,572	−28,863	−29,870
	(160)	(135)	(184)	(225)	(244)	(271)	(340)
33	−302	−327	−19,933	−29,401	−27,169	−27,875	−27,860
	(189)	(154)	(224)	(259)	(281)	(315)	(355)
Conventional Estimation							
Average	−2,419	−1,247	−14,626	−22,392	−20,379	−20,516	−20,741
	(51)	(50)	(45)	(45)	(49)	(49)	(51)

Notes: The table reports estimated average and age-at-birth-specific effects of motherhood on annual labor earnings of women. The estimation follows Equation (4). Standard errors clustered at the individual level are reported in parenthesis. The upper part of the table gives the average over all age-at-birth groups which is obtained by aggregating the age-at-birth-specific estimates in the middle part using each group's sample share as weights (as in Equation (5)). The lower part of the table gives the estimates from a conventional event study following Equation (3). The stacked DiD estimation is based on 4,478,828 observations, the conventional estimation on 2,747,794. The estimates tabulated here are plotted in Figure 3.
Source: Own calculations based on the SIAB (see Section II for a description).

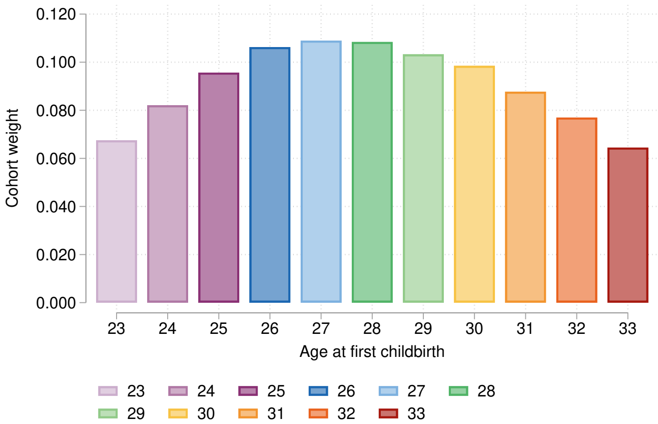


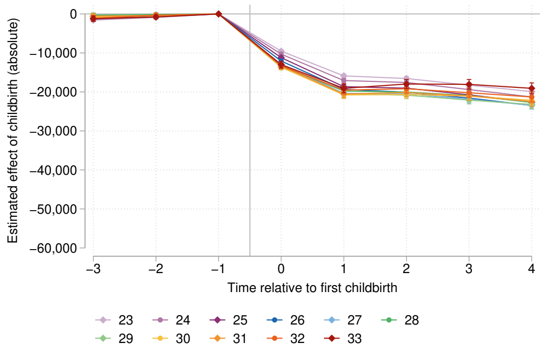
Figure A.6: Sample shares of age-at-birth groups (used as weights to calculate weighted-average estimates).

Notes: The figure plots the age-at-birth groups weights used to calculate the weighted-average effects after the first childbirth following Equation (5). They correspond to the share of each age-at-birth group in the sample, thus also indicate how the age at the first childbirth is distributed. The group-specific estimates are plotted in Figure 3a, the resulting average in Figure 3b.
Source: Own calculations based on the SIAB (see Section II for a description).

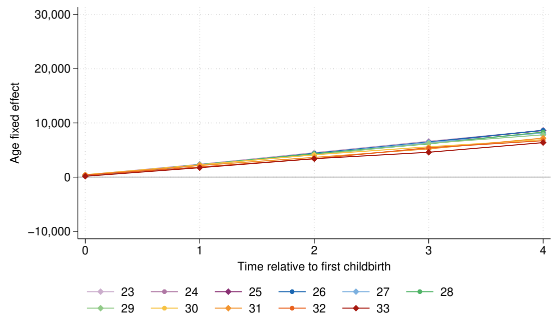
Table A.3: Average and age-at-birth-specific effects on annual earnings of mothers in relative terms.

	Time relative to 1st birth						
	−3	−2	0	1	2	3	4
Re-scaled with counterfactual							
Age-at-birth group							
23	−0.024	−0.012	−0.536	−0.819	−0.770	−0.770	−0.775
24	−0.002	0.003	−0.539	−0.830	−0.784	−0.787	−0.792
25	0.001	0.003	−0.545	−0.845	−0.793	−0.794	−0.801
26	0.007	0.006	−0.547	−0.845	−0.792	−0.791	−0.796
27	0.026	0.017	−0.555	−0.848	−0.788	−0.786	−0.790
28	0.030	0.016	−0.559	−0.847	−0.786	−0.781	−0.785
29	0.034	0.011	−0.558	−0.842	−0.775	−0.770	−0.770
30	0.017	0.008	−0.558	−0.840	−0.764	−0.757	−0.753
31	−0.001	0.001	−0.562	−0.824	−0.742	−0.738	−0.737
32	−0.004	0.004	−0.556	−0.807	−0.720	−0.717	−0.711
33	−0.010	−0.010	−0.561	−0.789	−0.694	−0.686	−0.670
Average	0.009	0.006	−0.553	−0.833	−0.768	−0.765	−0.766
Conventional estimate	−0.097	−0.048	−0.532	−0.789	−0.700	−0.689	−0.683
Re-scaled with pre-birth levels							
Age-at-birth group							
23	−0.013	−0.008	−0.483	−0.819	−0.842	−0.916	−1.001
24	−0.001	0.002	−0.498	−0.841	−0.867	−0.946	−1.032
25	0.001	0.002	−0.511	−0.868	−0.888	−0.968	−1.056
26	0.005	0.005	−0.520	−0.879	−0.901	−0.977	−1.062
27	0.019	0.014	−0.531	−0.892	−0.901	−0.972	−1.043
28	0.022	0.013	−0.548	−0.906	−0.910	−0.968	−1.038
29	0.025	0.010	−0.560	−0.918	−0.907	−0.960	−1.014
30	0.013	0.007	−0.575	−0.933	−0.902	−0.943	−0.982
31	−0.001	0.001	−0.594	−0.927	−0.882	−0.919	−0.968
32	−0.003	0.004	−0.587	−0.902	−0.845	−0.884	−0.915
33	−0.009	−0.010	−0.596	−0.878	−0.812	−0.833	−0.832
Average	0.007	0.005	−0.545	−0.890	−0.883	−0.941	−1.003
Conventional estimate	−0.092	−0.047	−0.554	−0.847	−0.771	−0.776	−0.785

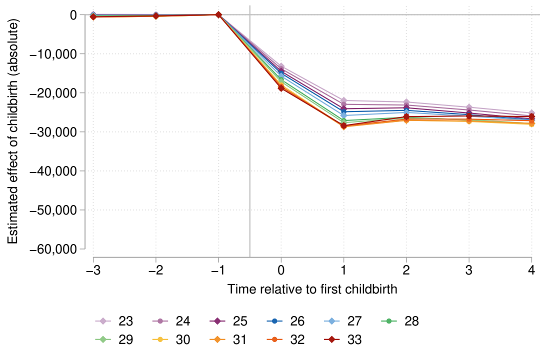
Notes: The table reports estimated average and age-at-birth-specific effects of motherhood on annual labor earnings of women in relative terms. The underlying estimation follows Equation (4), its results are tabulated in Table A.2. Estimates that are re-scaled with a counterfactual are divided by the prediction from the regression that omits the event time indicators; estimates that are re-scaled with the pre-birth outcome levels are divided by the earnings levels of each respective age-at-birth group at event time −1. Averages over all age-at-birth groups are obtained by aggregating the age-at-birth-specific estimates using each group's sample share as weights (as in Equation (5)). The conventional estimate is obtained following Equation (3). The estimates tabulated here are plotted in Figure 5.
Source: Own calculations based on the SIAB (see Section II for a description).



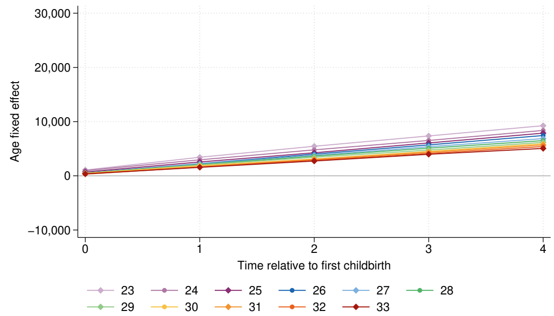
(a) No vocational training: effect of motherhood.



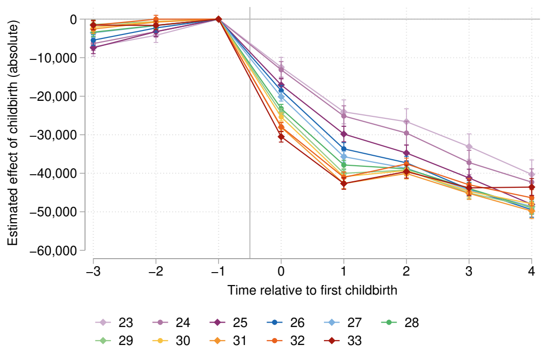
(b) No vocational training: counterfactual earnings.



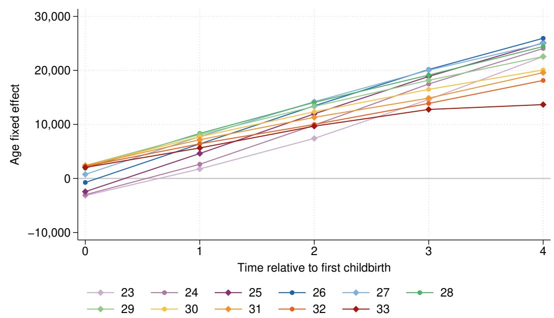
(c) Vocational training: effect of motherhood.



(d) Vocational training: counterfactual earnings.



(e) University: effect of motherhood.



(f) University: counterfactual earnings.

Figure A.7: Effects of motherhood and counterfactual outcomes by education prior to childbirth.

Notes: The left-hand side of the figures plots the estimated effects of motherhood (along with 95 percent confidence intervals) on annual labor earnings by education level. The estimations follow Equation 4. The right-hand side of the figure plots the education-specific age fixed effects taken from the respective estimations as measure of the earnings development of the control groups. Education is measured in the year prior to childbirth.
Source: Own calculations based on the SIAB (see Section II for a description).

Online Appendix

B The German Socio-Economic Panel (SOEP)

We complement the main analyses that use administrative data from the SIAB with survey data from the German Socio-Economic Panel (SOEP, Goebel et al. 2019). The SOEP is a well-established panel study that started in 1984 and surveys around 12,000 households and their members each year. Along with detailed socio-demographic information it provides data on labor force status, labor earnings, working hours, occupations as well as on the household context of mothers. Importantly, it also records the full birth histories that allow to identify mothers and when they have given birth. We apply similar selection criteria as for the SIAB data, i.e. we restrict to mothers who give their first childbirth in West Germany while being employed. We allow a higher maximum age at the first childbirth of 44 to capture the upper tail of its distribution. With these restrictions, our SOEP sample consists of data on 2,585 mothers for the period 1984 to 2020. As for the SIAB data, all monetary values are in real terms for the base year 2015.

Despite the different data sources and collection procedures, mothers in our samples from the SIAB and SOEP data show similar characteristics with respect to their age at the first childbirth and their labor market experience (see Table B.1). The comparison, however, shows that the SIAB understates mothers' education levels, especially for those with tertiary degrees. To a large part, this is likely not a sample selection issue, but due to a well-documented reporting problem (Fitzenberger, Osikominu, and Völter 2005). Mothers from the SOEP earn slightly more in the year prior to the first birth (€ 27,400 vs. € 26,900), which is in line with expectations as the SOEP includes individuals who are civil servants or self-employed, i.e. two groups who are typically higher up in the earnings distribution. If we exclude them from the sample, average pre-birth earnings in the SOEP are closer to the SIAB at € 26,700.

Table B.1: Comparison of summary statistics for mothers in the year prior to their first childbirth: SOEP and SIAB.

	Mean	SD	Min.	p25	p50	p75	Max.
Socio-Economic Panel							
Labor earnings	27,382	17,246	337	16,160	26,148	35,393	182,917
Education							
Share no vocational degree	0.08	0.28	0	0	0	0	1
Share vocational degree	0.68	0.47	0	0	1	1	1
Share university degree	0.24	0.43	0	0	0	0	1
Total years of education	12.87	2.76	7	11	12	15	18
Years in employment							
Full-time	6.17	4.52	0	2.5	5.42	9.08	25
Part-time	0.93	1.99	0	0	0	1	18
Age at first birth	28.81	4.82	20	25	29	32	44
N	2,585						
Sample of Integrated Labor Market Biographies							
Labor earnings	26,906	16,595	0	17,294	25,952	34,382	376,882
Education							
Share no vocational degree	0.15	0.36	0	0	0	0	1
Share vocational degree	0.68	0.47	0	0	1	1	1
Share university degree	0.11	0.32	0	0	0	0	1
Years in employment	6.01	3.98	0	2.90	5.45	8.58	22.22
Age at first birth	28.53	4.47	20	25	28	32	38
N	186,229						

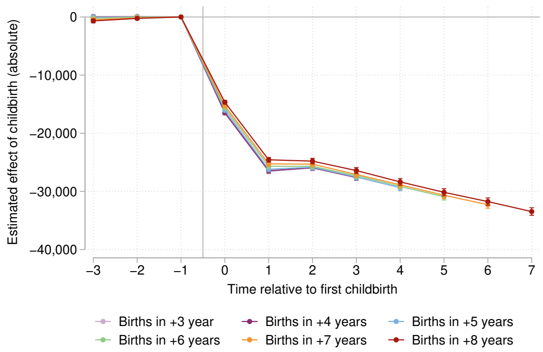
Notes: The table collects summary statistics of mothers who gave their first childbirth in West Germany while being employed. The SIAB data covers the period 1975–2021, the SOEP 1984–2020. Labor earnings are measured in Euro, deflated to the base year 2015. The SOEP data without conditioning on employment at childbirth show that 83 percent of mothers are in employment when they give birth. The education measure in the SIAB uses a variable provided by the Institute for Employment Research that accounts for the considerable share of missing values in the raw data using an imputation procedure based on Fitzenberger, Osikominu, and Völter (2005); a share of 6 percent of missing values nevertheless remains.
Source: Own calculations based on the SOEP and SIAB.

C Robustness Checks

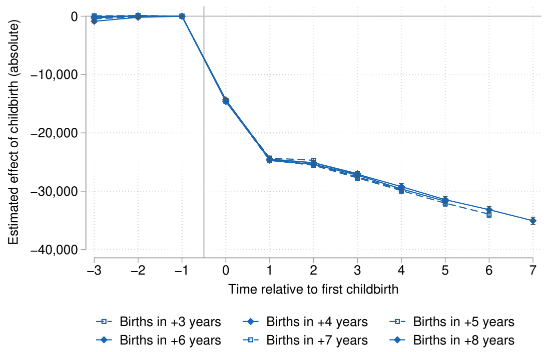
Table C.1: Robustness checks

	Time relative to 1st birth						
	−3	−2	0	1	2	3	4
I. Main specification (as in Table A.2)							
Stacked DiD Average	213	141	−16,114	−26,227	−25,931	−27,599	−29,345
	(113)	(87)	(127)	(154)	(172)	(193)	(236)
Conventional	−2,419	−1,247	−14,626	−22,392	−20,379	−20,516	−20,741
	(51)	(50)	(45)	(45)	(49)	(49)	(51)
II. Earnings > 99^{th} percentile dropped							
Stacked DiD Average	96	115	−15,721	−25,307	−24,702	−26,121	−27,606
	(100)	(73)	(109)	(135)	(150)	(165)	(195)
III. Mothers with censored earnings in $t = -1$ dropped							
Stacked DiD Average	−78	39	−15,726	−25,266	−24,666	−26,065	−27,539
	(102)	(75)	(111)	(137)	(153)	(170)	(200)
IV. Mothers who are unemployed after childbirth dropped							
Stacked DiD Average	258	245	−16,882	−26,714	−25,802	−27,502	−29,263
	(130)	(99)	(142)	(175)	(196)	(222)	(273)
V. Sample including mothers who are unemployed when giving birth							
Stacked DiD Average	648	535	−15,318	−24,502	−24,218	−25,691	−27,150
	(109)	(85)	(119)	(147)	(163)	(183)	(225)
Conventional	−1,768	−748	−13,961	−21,014	−19,200	−19,309	−19,501
	(48)	(48)	(43)	(42)	(46)	(46)	(47)

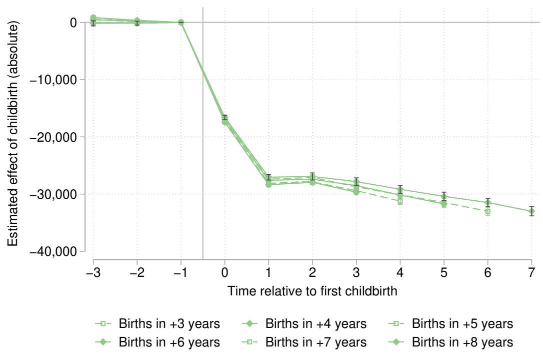
Notes: The table compares the average effects of motherhood estimated from our main specification (following Equation (4), listed in part I of the table) with those from several robustness checks. Standard errors clustered at the individual level are reported in parenthesis. Parts II and III repeat the estimation after dropping earnings above the 99^{th} percentile and mothers with censored earnings in the pre-birth year, respectively. For the estimates in part IV, mothers, for whom an unemployment spell of at least 30 days in at least one post-birth year is recorded, are dropped. Part V repeats the stacked DiD and the conventional estimation, but including also mothers who are recorded as unemployed when they give birth.
Source: Own calculations based on the SIAB (see Section II for a description).



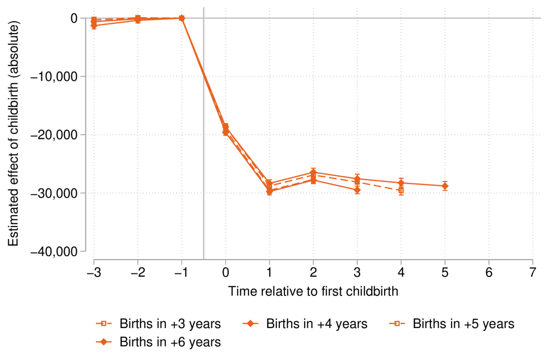
(a) Average results over all age-at-birth groups.



(b) Age-at-birth-specific results for mothers giving birth at 26.



(c) Age-at-birth-specific results for mothers giving birth at 29.



(d) Age-at-birth-specific results for mothers giving birth at 32.

Figure C.1: Estimates using mothers who give three up to eight years later in the control group.

Notes: The figure plots robustness checks for the main results that are plotted in Figure 3. It collects results from multiple estimations that use mothers who give birth from three to eight years later than the respective age-at-birth group as control units, whereas the main specification uses mothers who give birth up to five years later.
Source: Own calculations based on the SIAB (see Section II for a description).

D Analysis of the Conventional Re-Scaling Approach

A common step when estimating child penalties is to re-scale the event study estimates from levels to changes in percentages. Percentage changes are easier to interpret and allow straightforward comparisons of child penalties across different settings such as countries, points in time or policy regimes. The conventional approach introduced by Kleven, Landais, and Søgaard (2019) re-scales by calculating

$$P_l = \frac{\hat{\beta}_l}{\mathrm{E}\left[\tilde{Y}_{iat}|l\right]}, \tag{D.1}$$

where $\tilde{Y}_{iat}$ is the prediction from the previous regression (as in Equation (3)) when omitting the relative event-time dummies. $\mathrm{E}\left[\tilde{Y}_{iat}|l\right]$, the earnings levels that only depend on fixed effects for age and calendar year, are intended to proxy for earnings in a counterfactual state in which a woman does not have children. The child penalty P_l then gives the percentage difference between the earnings of mothers and counterfactual earnings of women without children.

This re-scaling step introduces an additional source of potential biases as it still contains treatment effects and makes comparisons between units that have been treated at different points in time. When predicting counterfactual earnings based on age and year fixed effects alone, the resulting $\tilde{Y}_{iat}$ is not restricted to only use specific observations and, therefore, consists of observations both from the pre- and the post-birth period. Since the share of women who already had their first child increases in age, the composition of the counterfactual in terms of including pre- or post-birth observations changes in age at childbirth as well. Figure D.1 illustrates this by plotting the share of pre-birth observations that is used in each relative time period from -5 to $+10$ to construct the counterfactual when estimating and re-scaling child penalties following Equations (3) and (D.1). Table D.1 provides the according numbers for five selected age-at-birth groups. It shows how substantial and quickly the composition of the counterfactual changes when moving from younger to older ages at first birth and further away from the year of the first birth. In the majority of the sample, the counterfactual for post-birth earnings is constructed from post-birth earnings which do not give an adequate depiction for a situation without having children. At the median age at first birth, 28, already the counterfactual for the last pre-birth year consists of more post- than pre-birth observations of other mothers. For mothers of age 35 (the 90^{th} percentile of age at first birth) the counterfactuals for the third and all following post-birth years include no pre-birth observations.

Abstracting from our example in Figure D.1 that is estimated using SIAB data, it is straightforward to formalize the relationship between age at first birth and the composition of counterfactual earnings. Assume, age at first birth A is distributed on some interval $\left[\underline{A}, \bar{A}\right]$. The – dataset specific – distribution function is $F_A(a) = P(A \leq a)$ and gives the share of observations that have given birth at or before age a. For any age at first birth a and any post-birth year t, $\sigma_{\text{pre}} = 1 - F_A(a+t)$ gives the share of remaining pre-birth observations in the sample that is not influenced by effects of previous childbirth. σ_{pre} is decreasing in age, such that that fewer and fewer suitable observations to construct counterfactual

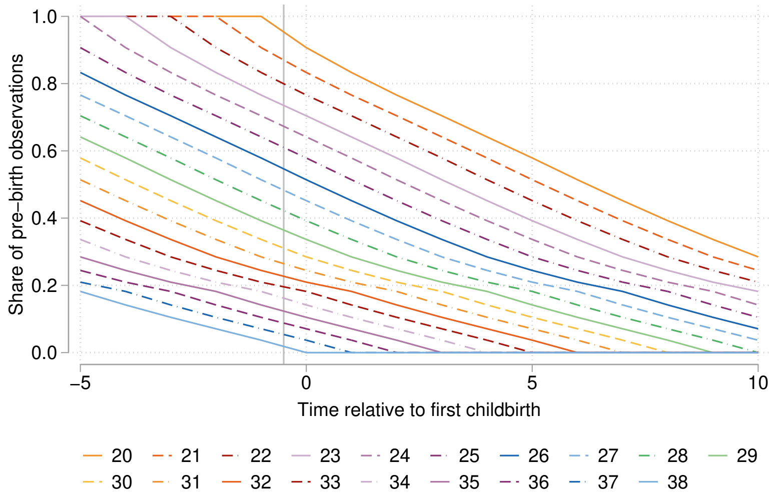


Figure D.1: Composition of counterfactual earnings $\mathrm{E}\left[\tilde{Y}_{iat}|t\right]$.

Notes: The figure plots the share of pre-birth observations in the counterfactual when estimating child penalties following Equations (3) and (D.1) by time around first childbirth and for each age-at-birth group.
Source: Own calculations based on the SIAB (see Section II for a description).

earnings are available when we look at older mothers. Therefore, this way to construct a counterfactual is not suitable to give an adequate depiction for a situation without treatment.

Table D.1: Share of pre-birth observations used to construct counterfactual earnings by relative time and age at first childbirth.

Years relative to first childbirth	Age at first childbirth				
	23	25	28	32	35
−5	1	0.907	0.704	0.452	0.285
−4	1	0.833	0.641	0.392	0.244
−3	0.907	0.766	0.579	0.337	0.210
−2	0.833	0.704	0.514	0.285	0.182
−1	0.766	0.641	0.452	0.244	0.142
0	0.704	0.579	0.392	0.210	0.105
1	0.641	0.514	0.337	0.182	0.071
2	0.579	0.452	0.285	0.142	0.037
3	0.514	0.392	0.244	0.105	0
4	0.452	0.337	0.210	0.071	0
5	0.392	0.285	0.182	0.037	0
6	0.337	0.244	0.142	0	0
7	0.285	0.210	0.105	0	0
8	0.244	0.182	0.071	0	0
9	0.210	0.142	0.037	0	0
10	0.182	0.105	0	0	0

Notes: The table reports the share of pre-birth observations that are available to construct counterfactual earnings as in Equation (D.1) by relative time around childbirth and for 5 levels of age at first childbirth. The age levels correspond to the 10th, 25th, 50th, 75th and 90th percentiles of the distribution of age at first childbirth.
Source: Own calculations based on the SIAB (see Section II for a description).