Reconciling Estimates of the Long-Term Earnings Effect of Fertility*

Simon Bensnes† Ingrid Huitfeldt‡ Edwin Leuven§

This version: October 14, 2025

Abstract

This paper reconciles different approaches to estimating the labor market effects of children. Combining elements from event-study and instrumental-variable estimators we find that while both approaches imply a 15 percent increase in the mother–partner earnings gap ("child penalty"), they differ in what drives this gap. The standard event study attributes it primarily to reduced maternal earnings, but our results suggest maternal changes account for less than half. We show that women time fertility as their earnings profile flattens, causing the event study to overestimate the maternal penalty. This finding has broader implications for event-study designs, as pre-trends may be uninformative about selection bias.

Keywords: Child penalty, female labor supply, event study, instrumental variable.

JEL codes: C36, J13, J16, J21, J22, J31.

*This paper has received funding from the Research Council of Norway (grant #256678 and #326391) and was previously circulated under the title "Event Studies, Endogenous Timing, and the Child Penalty." We thank Martin Andresen, Jon Fiva, James Heckman, Henrik Kleven, Magne Mogstad, Hessel Oosterbeek, and seminar participants at 15th IZA and 2nd IZA/CREST conference on Labor market policy evaluation (2019), OsloMet (2019), Statistics Norway (2019), Norwegian Institute for Public Health (2019), BI Norwegian Business School (2020), Zeuthen Workshop in Copenhagen (2021), University of Chicago (2022), and University of Oslo (2023) for comments.

†Frisch Centre.

‡BI Norwegian Business School and Statistics Norway.

§University of Oslo and Statistics Norway.

1 Introduction

Why do women earn less than men? Existing evidence finds that a substantial part of the gender pay gap can be attributed to the differential labor market costs of having children. While women's labor market earnings drop significantly around the time of their first childbirth, no such decline is apparent among men. This paper provides methodological and empirical contributions to this literature.

Estimating the impact of having children on labor market outcomes is a complex task, as fertility is intertwined with other factors that affect these outcomes. Neglecting these confounding factors results in omitted variable bias. To address this issue, the recent literature has mainly relied on event study approaches to study extensive margin fertility effects, pioneered by Korenman and Neumark (1992) and Waldfogel (1997), further developed by Anderson et al. (2003), Miller (2011) and Angelov et al. (2016) and more recently popularized by Kleven et al. (2019). These designs identify the effect of having children by comparing women who give birth at different times under (i) no anticipation and (ii) a common-levels/trends condition for untreated potential outcomes across timing cohorts—assumptions that can be violated if fertility timing is correlated with the slope of women's counterfactual earnings profiles.

An alternative approach proposed by Lundborg, Plug, and Rasmussen (2017) (henceforth referred to as LPR) is to use IVF (in vitro fertilization) as an instrumental variable for fertility. They showed that, given participation, the outcome of IVF treatment is conditionally as good as random, and therefore can be used to estimate the causal effect of fertility on earnings. To ensure identification, this approach (henceforth LPR-IV) requires the standard instrumental variable assumptions: conditional independence (given age, education), monotonicity, and exclusion (no direct effect of IVF success except through births).

The event-study and instrumental-variable approaches target different treatment effects. Event studies center time at birth and estimate dynamic treatment effects of fertility: the effect of having a child of a *given* age. In contrast, Lundborg et al. (2017) center time at the IVF attempt, and estimate the effect of having a child (of *any* age) at a given point in time since the attempt.

The two approaches also operate under distinct assumptions. This implies that if the exogeneity of fertility timing fails then the event study is compromised. The assumptions of the IV approach are challenged by delayed fertility–women who fail a first IVF trial catching up later–as this not only changes the complier group over time, but also the age composition of the children. As pointed out by Lundborg et al. (2017), the fertility response is underestimated (a positive bias) if the impact

of having children on female labor earnings is particularly large when children are young.

In a first contribution we explore the differences between these two models, and show that the estimated treatment effects in LPR-IV can be viewed as latent mixtures of different dynamic (age-of-child specific) treatment effects. We find that, as compliance to the instrument falls with time since the attempt, the LPR-IV model assigns increasingly more negative weight on the effect of children born after the first IVF trial because these are overrepresented in the control group. This mapping also delivers a hybrid IV approach (henceforth referred to as event-IV) which exploits the design of Lundborg et al. (2017), but estimates age-of-child specific treatment effects that are comparable to the event-study, while addressing the complications stemming from delayed fertility. We also propose a simple increasing-horizon homogeneity check: we re-estimate age-of-child effects on progressively larger balanced horizons and test for stability; in our data, estimates are highly stable.

Second, we bring these three estimators–event, LPR-IV, and event-IV–to the data, and re-examine the labor market effects of having children using administrative data on IVF treatments, family links and labor market outcomes for the entire Norwegian population. Across all models we observe a considerable 15 percent increase in the long-term earnings gap between parents, often referred to as the *child penalty*. The models disagree however on the extent to which women or partners are driving this gap.

The policy implications of child penalties rest on whether it is the woman or the partner who drives this result. If the impact of children on parental earnings gaps is caused by partners earning more while women's earnings remain unchanged, then policies aimed at promoting female labor supply, such as flexible work arrangements, may not be effective in closing the gap. Conversely, if gaps result from a reduction in women's earnings, such policies may work as intended.

The event-study model suggests that for earnings nearly all of the child penalty is driven by women, and points to large negative long-run effects on maternal earnings of around 13 percent, in line with previous event-study estimates from other Scandinavian countries, including Norway (e.g., Kleven et al., 2019; Andresen and Nix, 2022). In contrast, the LPR-IV model reveals negligible point estimates for women, in line with recent evidence from Denmark which can estimate fertility effects on the very long run (up to 24 years after birth) (Lundborg et al., 2024). Finally, the event-IV model falls in between, and attributes about half of the child penalty to women and finds a reduction of 7 percent. Turning to partners' earnings, the ordering of the estimates goes therefore in the opposite direction: The event-

study model estimates an increase of 2 percent, while the LPR-IV model estimates an increase of 16 percent. The event-IV model once again falls in between at an increase of around 9 percent.

Focusing on the results from the event-IV model, we find that the earnings response for women is primarily driven by changes in employment status. For partners we also see employment responses, but the effect on earnings is also partially explained by both responses on hours worked and hourly wages.

Third, we explore the sources of bias and differences in estimates between models. The positive bias of the LPR-IV model is explained by increasingly more negative weight on the negative child-age effects for younger children. By the end of our estimation period the combined absolute size of these weights is about three-thirds compared to the weight that the contemporaneous child-age effect receives.

The difference between the standard event-study model and the event-IV estimates of the effect on women's long-run earnings is largely accounted for once we adjust earnings profiles for time since the IVF trial (a predetermined variable). The remaining difference between the event-study estimates with these timing controls and our event-IV estimates is driven by the always takers to our instrument–women who conceive naturally, adopt, or are successful at later trials–having higher earnings.

Fourth, even without relying on instrumental variable assumptions, we therefore find that seemingly robust findings can substantially change when controlling for the typically unobserved timing of women's fertility attempts. We estimate the counterfactual earnings profiles to understand how endogenous timing of fertility biases estimates from the standard event-study setup. These results reveal that women have their first child when their earnings profiles start to flatten out, and that women who have children later are on wage profiles that continue to grow beyond those of women who have children earlier. This is clear evidence of a violation of the parallel-trend assumption.

Finally, the type of selection we uncover not only show that pre-trends are not always informative of violations of parallel trends in the treatment period, but also that extrapolations of the pre-trend–a common robustness check formalized in Rambachan and Roth, 2023)–may exacerbate the bias relative to the standard-event study specification. Pre-trends (and bias) may also arise because of confounding treatment effect heterogeneity as discussed in a series of recent advancements (see, e.g. Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; Borusyak et al., 2024; Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020). However, we

find that while estimates based on Callaway and Sant'Anna (2021) align pre-trends, they exhibit a more negative bias taking the event-study estimates that adjust for endogenous timing or the event-IV estimates as a benchmark.

Our results speak to women who conceive through IVF. In Norway about six percent does so, and they therefore represent a non-negligible (and increasing share) of the broader population. It is however an open question to what extent our results generalized elsewhere. We find that IVF women are similar to non-IVF women. Women have slightly more education and conceive at slightly higher ages, but are otherwise similar to the general population of women.

We also interpret our results as suggesting that IVF births are not substantially different from regular births. A common concern is "disappointment" after an unsuccessful IVF attempt, which may affect behavior or mental health (e.g. Gallen et al., 2023; Martinenghi and Naghsh-Nejad, 2025). However, disappointment from not conceiving is unlikely to be specific to IVF, and standard event-study analyses for the broader population also likely include disappointed women in the comparison group.[1] This is especially true for our event-study estimates for IVF women, where any disappointment effect, if present, is also embedded in the counterfactual. Together, these considerations suggest that disappointment effects are unlikely to explain the major differences across models. Finally, while we document mental-health impacts, they are short-lived and affect a minority of IVF women, and our estimated effects of children are virtually unchanged when we control for these channels or exclude potentially affected women. This pattern is inconsistent with large average effects operating through this channel.

In addition to the literature cited above, this study also relates to a longstanding literature on the relationship between fertility and female labor supply. Early dynamic labor supply models incorporated fertility decisions by including child care costs in the index function of dynamic choice models (see, e.g. Adda et al., 2017; Heckman and McCurdy, 1980; Hotz and Miller, 1988). Recognizing the endogeneity of fertility, a strand of papers has used information on e.g. contraceptives, infertility shocks, and miscarriages to estimate the impact of fertility on labor supply (see, e.g. Hotz et al., 2005; Cristia, 2008; Aguero and Marks, 2008; Miller, 2011; Gallen et al., 2023). The endogeneity concern has also been addressed with twin-birth and same-sex instruments, though these are only suitable to study effects along the intensive fertility margin (e.g. Rosenzweig and Wolpin, 1980; Angrist and Evans, 1998; Bronars and Grogger, 1994).

[1] Similar issues arise in other IV settings where eligibility rules or lotteries may generate disappointment among nonrecipients—for example, school assignment, housing assistance, job training, or health-care access.

In the next section we start by providing the relevant institutional background information concerning IVF treatments as well as the social benefit system that may mediate the impact of motherhood on labor market outcomes. Section 3 describes the registry data and sample construction. We then present the existing estimators in Section 4, and connect them to the new empirical approach of this paper. Section 5 investigates the validity of success in the IVF trial as an instrumental variable. The child-penalty estimates are reported and discussed in Section 6, after which Section 7 traces the sources of their differences and discusses the resulting bias. We consider the external validity of our findings in Section 8. Section 9 summarizes and concludes our analysis.

2 Institutional Context

IVF

In vitro fertilization (IVF) is a method for women to become pregnant after failing to conceive through regular intercourse. The process is initiated by intake of medicines designed to increase the number of eggs the patient normally produces during ovulation. The eggs are then collected and manually fertilized with donor sperm or sperm from the woman's partner at a clinic. The fertilized egg (zygote) is then cultured for 2-6 days in a growth medium. Once an egg is successfully fertilized it can be implanted in the woman's uterus. The default IVF procedure during our period of observation was a single embryo transfer. This means that IVF had a relatively low occurrence of multiple births (Bhalotra et al., 2019; Bhalotra and Clarke, 2019) (in our data, <0.1).

The receipt of IVF treatment in Norway is regulated by the Biotechnology Law. Women who fulfill the following eligibility criteria are entitled to three treatments at a public hospital: (i) infertility diagnosis certified by a physician, which requires a failure to conceive after a year of regular intercourse; and (ii) live in a marriage-like relationship. A treatment includes both harvesting of eggs and implantation of fertilized eggs. In cases where multiple eggs are fertilized and frozen after one retrieval, the implantation of these eggs are considered part of a single treatment. It is therefore possible to go through several rounds of inserting fertilized eggs within one treatment. In our analyses we refer to trials or attempts as the *insertion* of eggs, which is identified in the data since hospitals are reimbursed by the government for these procedures. Public institutions prioritize childless couples where the age of the women is below 39 and her BMI is below $33 kg/m^2$.

The co-payment for the first three treatments at a public hospital is about NOK

6 000 (USD 670 in 2019) per treatment, and covers medicines and pharmaceutical expenses. Private institutions offer an alternative to public hospitals and comprise 15-20% of the market. Private options are considerably more expensive – around NOK 100 000 (USD 10,900) for a single treatment – but may have shorter wait times and more flexibility in terms of age requirements.

Social benefits

The relationship between fertility and labor market outcomes may be shaped by labor market institutions and social insurance systems. Since the 1970s, Norway has implemented comprehensive parental support policies (NOU 2017:6, 2017). During our study period, parents were entitled to approximately one year of parental leave after childbirth, with two options: either slightly less than a year at 100% wage replacement or a longer period (extended by ten weeks) at 80% wage replacement.

The support system extends beyond parental leave. Pregnant women can request welfare support if their working conditions pose potential risks to maternal or fetal health. Legal protections prohibit employers from pregnancy-based discrimination in hiring, promotion, and termination decisions. The system also provides generous sick leave benefits, allowing workers to take time off both for personal illness and to care for sick children. Furthermore, beginning in the early 2000s, the national government significantly expanded formal childcare, making subsidized facilities widely accessible to virtually all families (Andresen and Havnes, 2019; Drange and Havnes, 2019). Together, these various support mechanisms—spanning pregnancy, childbirth, and child-rearing periods—may help offset potential fertility-related earnings losses.

3 Data sources and sample

Data and variables

The empirical analysis is based on data that combine several administrative registers from Statistics Norway and the Norwegian Directorate of Health. Every Norwegian resident receives a unique personal identifier at birth or upon immigration, enabling us to match the health records with administrative data for the entire resident population of Norway, which contains information on birth and death dates, sex, district and municipality of residence, country of origin, and education. The data further include family links, allowing us to match women with their partners and children. These data are available for us up until 2022.

Every IVF treatment administered at a public hospital is recorded in the Norwegian Patient Registry. This registry contains complete patient level observations of all visits financed by the Norwegian public health care system. From 2008 onward, the records contain patient identifiers that can be linked to administrative data. The patient data include information on primary and secondary diagnoses (ICD10), surgical/medical procedures (NCSP/NCMP)[2] , exact time, date and place of admissions and discharges. We use these data to identify IVF trials from the surgical procedure code "LCA 30 - Transfer of zygote or embryo to uterus in assisted fertilization." Additionally, we construct a variable with counts of the number of days spent (both inpatient and outpatient stays) at the hospital in a given year. These data are available over the period 2008 to 2017.

In addition to health records from hospital visits, we retrieve data on visits to primary care physicians from the Control and Payment of Health Reimbursement (KUHR). These data include the date of visit, diagnosis codes (ICPC-2) and reimbursement fees. From these data, we construct a variable measuring the number of GP visits in a given year, and a binary indicator for whether at least one visit was coded with a psychological symptom or disorder. This indicator includes both milder symptoms, such as anxiety, stress, or sleep problems, and more severe diagnoses, such as depression, psychosis, or substance abuse. We also define a separate indicator restricted to the severe diagnoses only (see Appendix Table A12 for the full list of codes). The data are available for us from 2006 to 2017.

Our main labor market outcomes are derived from the employer-employee registry. This registry contains information on start and stop dates of a job spell, as well as the corresponding labor income, occupation, sector and contracted hours. We have access to these data for the period 2004 to 2022.

We define four variables to capture individuals' labor market attachment. Our main outcome, *Earnings*, captures yearly labor income, excluding parental leave benefits. We adjust for inflation using 2015 as the base year. *Employed* is a binary indicator equal to one if the individual has labor income more than the substantial gainful activity level in a given year, zero otherwise.[3] *Hours* is the number of contracted hours over a year, and *Hourly earnings* is the wage rate, calculated by dividing earnings by hours. In the main part of the paper we focus on the effects on labor earnings and report estimates for the other outcomes in the appendix. We

[2]Standardized systems developed for consistent coding of surgical and medical procedures across Nordic countries.

[3]The substantial gainful activity level ("basic amount") was equivalent to NOK 90 068 in 2015. The basic amount is used by the Norwegian Social Insurance Scheme to determine eligibility for welfare benefits.

also report estimates for earnings *including* all social benefits in Appendix A.8.

Sample

Our main analysis sample consists of 10,033 women who had at least one IVF trial over the period 2009 to 2016, and who did not have any children prior to their first attempt. We exclude women with any IVF trial in 2008, which is the first year in which IVF treatment can be identified in our data. As most women pursue a second attempt within twelve months upon failure at first attempt, this allows us to restrict our sample to women who receive IVF treatment for the first time. We also restrict the sample to women who are at least 18 years old, and who were registered with a partner in the year of the first IVF treatment.[4] For comparison, we also construct a sample of women who had children without IVF treatment. This sample consists of women who had their first child in the same period as the successful IVF women (2009 to 2017), and who were registered with a partner in the year of conception.

Descriptive statistics are presented in Table 1. We follow Lundborg et al. (2017) and define attempts as successful if (i) the woman gives birth within five to ten months of the trial, and (ii) there were no other trials in the time between the trial and the birth. In our sample, the average number of IVF trials is about 2.8, the success rate after one trial is 31 percent, and the end-of-period success rate is 63 percent. In total 83 percent of the IVF women eventually have at least one child. The difference between realized fertility and IVF success at the end of the sample period is explained by child birth without the aid of IVF, adoption, and possibly also children born after successful IVF attempts at private clinics. At the end of our observation period, 30 percent of the IVF women have one child, and 42 percent have two children, 9 percent have three children, and virtually none have four children or more. Among the IVF *mothers*, i.e. those who have at least one child, 36 percent (0.30 / 0.83) have one child, 53 percent have at least two children, and 12 percent have three or more children (see also Table A2). For comparison, non-IVF mothers are more likely to have two or more children, 23 percent have one child, while 60 percent have two children, and 17 percent have three or more children.[5]

The average age at first trial is just below 32, while non-IVF women have their first child at age 28. The education level is very similar but slightly higher for IVF

[4]Only women in stable unions are eligible for public IVF treatment. However, this does not require a formal marriage, and partnership may therefore not show up in the administrative data. When restricting our sample to women with a registered partner, we lose 14 percent of the IVF participants, and 46 percent of the non-IVF women.

[5]When we limit our sample to the women we can observe for 3 years after the trial, 62% of those who failed their first trial have at least one child, 74% of women have one child (regardless of outcome of first trial).

Table 1. Descriptive statistics for IVF women and non-IVF women

	(1) IVF	(2) Non-IVF	(3) Difference	
Woman characteristics				
Number of IVF attempts	2.84			
Success, first trial	0.31			
Success, endpoint	0.63			
Fertility, endpoint	0.83	1		
Total number of children	1.47	1.97	-0.50	(0.01)
0 children	0.17	0		
1 child	0.30	0.23	0.07	(0.00)
2 children	0.44	0.60	-0.15	(0.01)
3 children	0.09	0.16	-0.07	(0.00)
4 children	0.01	0.02	-0.01	(0.00)
Age	31.8	28.4	3.41	(0.05)
Education				
- Compulsory	0.14	0.17	-0.03	(0.00)
- High School	0.24	0.23	0.01	(0.00)
- Bachelor	0.42	0.41	0.01	(0.01)
- Master	0.20	0.19	0.01	(0.00)
Earnings (1000 NOK)	362.7	289.9	77.2	(1.86)
Hours (FTE)	0.88	0.79	0.10	(0.00)
Employed	0.80	0.67	0.14	(0.00)
Hourly earnings (NOK)	221.1	197.5	23.6	(1.85)
Sickness absence days	15.0	11.1	4.85	(0.30)
Visits to general practitioner (GP)	2.51	2.16	0.50	(0.02)
Any (mild/severe) psychological symptoms	0.14	0.12	0.02	(0.00)
Severe psychological symptoms	0.06	0.06	0.00	(0.00)
Hospital days	2.13	1.01	1.25	(0.04)
Partner characteristics				
Age	35.1	31.2	3.9	(0.06)
Female	0.01	0.01	0.00	(0.00)
Education				
- Compulsory	0.17	0.20	-0.03	(0.00)
- High School	0.39	0.37	0.02	(0.01)
- Bachelor	0.27	0.26	0.01	(0.00)
- Master	0.17	0.17	0.00	(0.00)
Earnings (1000 NOK)	454.9	385.4	69.5	(2.90)
Hours (FTE)	0.84	0.78	0.06	(0.00)
Employed	0.84	0.76	0.08	(0.00)
Hourly earnings (NOK)	281.2	254.6	26.6	(2.42)
N Women	10 033	109 791		

Notes: Column (1) shows descriptive statistics for women who had at least one IVF trial over the period 2009 to 2016. Column (2) shows descriptive statistics for women who had at their first child without IVF treatment during the period 2009 to 2017. By construction, this includes only women who have at least one child. Column (3) shows the difference and corresponding standard error. Labor market outcomes and health indicators are measured as averages over the four years prior to the first IVF trial, or, for non-IVF women, prior to the approximate conception date. Education is measured in the calendar year before the IVF attempt / approximate conception date. Age is defined as the maternal age at the date of the IVF attempt / approximate conception date.

women, with 38 percent with high school education or lower, and 62 percent with a bachelor's degree or higher, compared to 40 percent and 60 percent for non-IVF women. While IVF women's average pre-trial earnings were 363,000 NOK (ca. 36,300 USD), non-IVF women earned 290,000 NOK per year. Among IVF women, 80 percent were employed, on average they worked the equivalent of 88 percent of a full-time position (FTE) per year, and earned 221 NOK per hour worked. For non-IVF women, 85 percent were employed, and their number of hours worked per year equaled 0.79 FTEs on average, yielding 198 NOK in hourly wages.

IVF women had somewhat higher utilization of health care services. Their pre-treatment sickness absence was 15 days per year, compared to 11 for non-IVF women; and they spent on average 2.1 days per year at a hospital, compared to 1 day for non-IVF women. The average number of visits to the GP was about 2.5 per year for IVF women, and 2.2 for non-IVF women. There was only a small difference in the share with at least one GP visit for any psychological symptom (severe or mild): 0.14 among IVF women and 0.12 among non-IVF women. In both groups, 0.06 had at least one visit related to severe psychological symptoms.

The average age of partners is 35 for IVF-women, compared to 31 for non-IVF women. The share registered with a female partner is one percent in both samples. The education levels of partners seem to be fairly similar across the two samples, with 27 percent holding a bachelor and 17 percent holding a master in the IVF sample, compared to 26 percent and 17 percent, respectively, in the non-IVF sample. Partners of IVF women earned on average 455,000 NOK and worked 0.84 FTEs per year, while partners of non-IVF women earned 385,000 NOK and worked 0.78 FTEs per year.

Compared to non-IVF women giving birth during the same period, we therefore see that IVF women tend to be somewhat older, and earn and work more, while their educational attainments are only marginally higher. The same patterns are also seen for their partners. In terms of our health measures the women are comparable, and while non-IVF women are more likely to have more than one child, their final fertility patterns are overall quite similar.

4 Estimating the effects of fertility on labor market outcomes

This section lays out our empirical framework in four steps. We begin with the birth-centered event-study design that is standard in this literature, detailing the identifying assumptions and the dynamic, age-of-child effects it recovers. We then review the IVF-based instrumental-variables approach of Lundborg et al. (2017)

(LPR-IV), which replaces the event-study exogeneity requirement with conditional randomness of IVF success and identifies extensive-margin fertility effects. Next, we formally connect the two approaches by showing that the LPR-IV estimand at a given time since the IVF attempt can be expressed as a mixture of age-of-child effects, clarifying why estimates may diverge when timing and age effects interact. Building on this link, we introduce an instrumental-variable event study ("event-IV") that centers time at birth, adjusts flexibly for the timing of the IVF attempt, and uses IVF success to address endogenous fertility timing while preserving dynamic, age-specific effects, and detail an over-identification check. We conclude by describing how we scale effects and define the child penalty to facilitate comparison with the existing literature.

4.1 Event study

To estimate how fertility affects women's labor supply we start by implementing the event-study specification that is standard in the literature and which centers time on the event of interest, T_i, the time of birth of woman i's first child. Given the age of the child in period t, $a_{it} \equiv t - T_i$, we can define child-age specific indicators $\mathbf{1}\{a_{it} = a\}$.

If we consider potential outcomes y_{it}^a for woman i in period t, then observed outcomes map to potential outcomes as follows

$$y_{it} = \sum_a (y_{it}^a - y_{it}^\infty)\mathbf{1}\{a_{it} = a\} + y_{it}^\infty \tag{1}$$

where superscript $a = \infty$ indicates the counterfactual of never having a child, and $a < \infty$ the counterfactual of having a child of age a (negative values of a refer to time before birth).

Most studies estimate equation (1) on samples of mothers, and make functional form assumptions that result in the following baseline event-study specification that we consider in the analysis below

$$y_{it} = \sum_{a \neq -1} \delta_a \mathbf{1}\{a_{it} = a\} + x_{it}'\phi + \tau_t + \epsilon_{it} \tag{2}$$

This specification anchors the counterfactual wage profile to a year prior to birth ($a = -1$). The coefficients on the child-age dummies, δ_a, allow for age-of-child specific effects on the outcome y_{it}. The counterfactual wage profile consists of controls x_{it} which adjust flexibly for women's age using dummy variables, and calendar year dummies τ_t. The main outcome and summary measure of women's

labor supply that we consider is yearly earnings from employment, but we also look at additional outcomes in section A.6.

If there are no anticipation effects, $y_{it}^{a} = y_{it}^{\infty}$ for all $a < 0$, and if untreated potential outcomes satisfy a common-levels condition,

$$E[y_{it}^{\infty} \mid T_i, x_{it}, \tau_t] = E[y_{it}^{\infty} \mid x_{it}, \tau_t], \tag{3}$$

then for each timing-cohort $T_i = g$ the cohort-specific average treatment effect (ATT) at event time a,

$$\delta_{ag} \equiv E[y_{it}^{a} - y_{it}^{\infty} \,\big|\, T_i = g,\ a_{it} = a],$$

can be identified using not-yet-treated units.[6] When the estimation restricts attention to mothers, the counterfactual outcome profile is therefore identified from pre-birth (untreated) outcomes, coming from differential timing of first birth across women.

In the event-study specification (2), which imposes a common set of event-time coefficients across timing cohorts, the OLS estimates of δ_a equal the simple ATT at event time a only under a cohort-homogeneity restriction $\delta_{ag} = \delta_a$ for all g; without homogeneity, the estimates of δ_a are (possibly non-convex) weighted averages of the δ_{ag}. Below we also report results using the event-study estimator proposed by Callaway and Sant'Anna (2021) (Figure A15) that relaxes the cohort-homogeneity assumption.

With few exceptions the baseline event-study specification in the literature that estimates the effect of children on mothers' labor supply does not include mother fixed-effects (e.g. Kleven et al., 2019). Adding mother fixed effects absorbs all time-invariant earnings heterogeneity, weakening the exogeneity assumption (3) to a common-trend assumption:

$$E[\Delta y_{it}^{\infty} \mid T_i, x_{it}] = E[\Delta y_{it}^{\infty} \mid x_{it}]$$

Below we also report event-study estimates that adjust for mother fixed effects (reported in Figure A16 in the appendix) for completeness.

Much of the literature also examines the earnings gap between women and men rather than focusing on women's earnings alone (e.g. Kleven et al., 2019; Andresen and Nix, 2022). Rather than assuming that women's earnings trajectories would be identical regardless of when they have children, this assumes that the earnings gap between women and men would evolve similarly in the absence of children. This approach comes of course at a cost: it only allows one to examine the difference in

[6] See f.e. de Chaisemartin and D'Haultfœuille (2020); Callaway and Sant'Anna (2021); Sun and Abraham (2021); Goodman-Bacon (2021); Borusyak et al. (2024)

outcomes between women and men, without being able to identify their separate impacts. We further discuss this distinction in Section 4.5.

4.2 Fertility effects of IVF (LPR-IV)

If the timing of birth correlates with unobserved levels and/or trends in earnings (e.g., women with lower or declining unobserved earnings potential tend to give birth earlier), the exogeneity condition of the event study fails and δ_a may be biased. This is why some have advocated for approaches that exploit arguably exogenous variation in fertility at the extensive margin.

One such alternative approach to identify fertility effects which does not rely on the event-study exogeneity assumption, is due to Lundborg et al. (2017), who argue that IVF births provide variation in fertility that is *conditionally* exogenous. Specifically, conditional on a woman's age, education, and the timing of IVF, the outcome of the first IVF attempt is assumed random.

Let IVF_i be the calendar year of woman i's first IVF treatment and re-index time by years since IVF: $p = t - IVF_i \in \{0, 1, 2, \ldots\}$, so $y_{ip} \equiv y_{i,\, t=IVF_i+p}$, etc. Define the instrument success$_i \in \{0, 1\}$, equal to one if the first IVF leads to a birth roughly nine months later. The "treatment" at horizon p is the extensive-margin indicator

$$\text{Fertility}_{ip} \equiv \mathbf{1}\{T_i - IVF_i \leq p\}.$$

Lundborg et al. (2017) estimate, separately for each horizon p, the following equation

$$y_{ip} = \gamma_p \, \text{Fertility}_{ip} + x_{ip}'\psi + \eta_{IVF_i} + \theta_p + u_{ip}, \tag{4}$$

where γ_p is the extensive-margin fertility effect of interest. To account for the endogeneity of fertility they instrument Fertility$_{ip}$ with IVF success$_i$:

$$\text{Fertility}_{ip} = \pi_p \, \text{success}_i + x_{ip}'\lambda + \zeta_{IVF_i} + \xi_p + \omega_{ip}. \tag{5}$$

Here x_{ip} includes flexible dummies for (woman's) age, and both equations control for IVF year, IVF_i (i.e. η_{IVF_i} and ζ_{IVF_i}) , and time since IVF, p (θ_p and ξ_p). In addition all controls are interacted with education to support conditional independence.[7] We refer to the 2SLS estimate of γ_p as the LPR-IV estimator.

Identification of γ_p at horizon p relies on: (i) instrument relevance ($\pi_p \neq 0$), (ii)

[7] While Lundborg et al. (2017) estimate (4) separately for each horizon p, we impose homogeneity in (4), which yields a more parsimonious mapping as shown below. This does not matter for the results. Figure A18 compares the p-stratified estimates with those from the specification in (4), and shows that these are nearly identical.

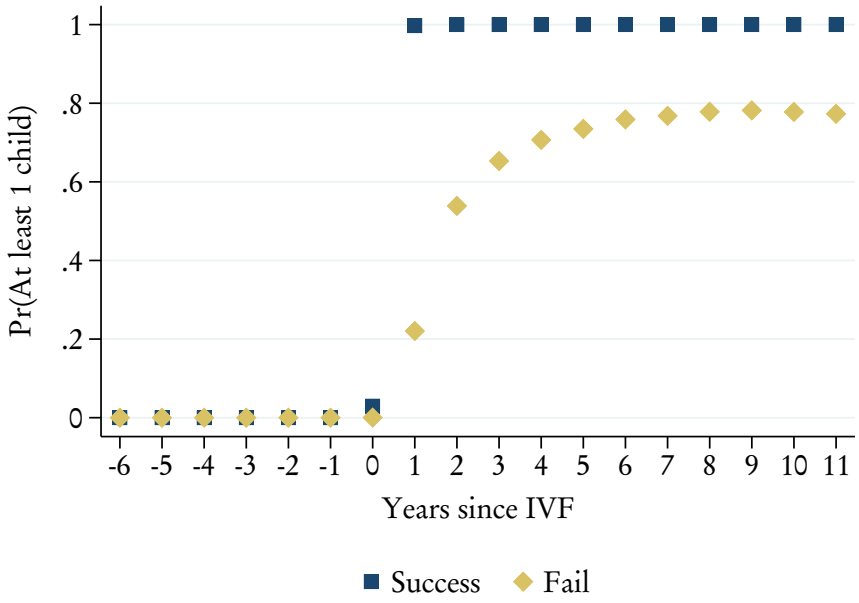


Figure 1. Fertility by success at first IVF trial

Note: Share of women having at least one child by year relative to first IVF treatment, grouped by success in first trial. The sample includes all women who underwent IVF treatment in 2009 to 2016, had no children prior to their first IVF attempt, and were registered with a partner at the time of the attempt (unique women = 10,033; observations = 173,480).

conditional independence of success_i given age, education, and IVF_i, (iii) monotonicity (mechanical here since $\text{success}_i = 1$ implies a birth by $p = 0$), and (iv) an exclusion restriction stating that, for $p \geq 1$, IVF success affects y_{ip} only through Fertility_{ip}. Under these conditions, the 2SLS coefficient $\hat{\gamma}_p$ has the usual Wald interpretation as an estimate of the local average treatment effect of fertility for women who have a child p years after the IVF attempt if the IVF is successful but not otherwise.

However, Lundborg et al. (2017) point out that this interpretation is compromised if IVF success not only affects whether a woman has a child in period p but also the age of that child. Figure 1 shows that delayed fertility is indeed a concern: many women with an unsuccessful first IVF subsequently conceive later. As a result, the first stage equals one on the very short run after nine months, (i.e. all women are very short-run compliers), and weakens as p increases (growing shares of always-takers by horizon p). This means that success_i not only changes whether a woman has had a child by p but also by the child's age at p. As a consequence, γ_p at longer horizons can mix extensive-margin effects with timing (age-of-child) effects and may understate short-run impacts when effects diminish with child age.

It is therefore not immediate how the IV estimates of γ_p compare to the dynamic,

child-age effects of the event study. We address this by formally linking γ_p to first-stage–weighted averages of child-age-specific effects in the next subsection.

4.3 Mapping LPR-IV fertility effects (γ_p) to child-age effects (δ_a)

Write the child's age at horizon p as $a_{ip} \equiv p - P_i$ where $P_i = T_i - IVF_i$ is the realized time-to-birth (with $a_{ip} = \infty$ if $P_i > p$). For each $p \geq 0$ and $a \in \{0, \ldots, p\}$, there is an age–horizon first stage

$$\mathbf{1}\{a_{ip} = a\} = \pi_{ap}\,\text{success}_i + x_{ip}'\lambda_a + \zeta_{a,IVF_i} + \xi_{ap} + \omega_{aip}, \tag{6}$$

Substituting these age-horizon first stages into the following identity[8]

$$\text{Fertility}_{ip} \equiv \sum_{a=0}^{p} \mathbf{1}\{a_{ip} = a\},$$

gives the first-stage (5) of LPR above where the first-stage coefficient π_p maps to the age-horizon specific first-stage coefficients as follows[9]

$$\pi_p = \sum_{a=0}^{p} \pi_{ap}$$

Similarly, write the horizon-p outcome as

$$y_{ip} = \sum_{a=0}^{p} \delta_{ap}\,\mathbf{1}\{a_{ip} = a\} + x_{ip}'\psi + \eta_{IVF_i} + \theta_p + u_{ip}, \tag{7}$$

then substituting the age-horizon first-stages gives the following reduced form

$$y_{ip} = \rho_p\,\text{success}_i + x_{ip}'\tilde{\phi} + \tilde{\eta}_{IVF_i} + \tilde{\xi}_p + \tilde{u}_{ip} \tag{8}$$

where the reduced-form coefficient ρ_p equals

$$\rho_p = \sum_{a=0}^{p} \delta_{ap}\,\pi_{ap}$$

[8] Note $\mathbf{1}\{a_{ip} = a\} = 0$ for all $a > p$.

[9] In addition, $\lambda = \sum_{a=0}^{p} \lambda_a$, $\zeta_{IVF_i} = \sum_{a=0}^{p} \zeta_{a,IVF_i}$, $\xi_p = \sum_{a=0}^{p} \xi_{ap}$, $\omega_{ip} = \sum_{a=0}^{p} \omega_{aip}$.

The fertility effect from the LPR-IV model, $\gamma_p = \rho_p/\pi_p$, is therefore a first-stage–weighted average of the age-horizon effects at p:

$$\gamma_p = \frac{\rho_p}{\pi_p} = \frac{\sum_{a=0}^{p} \pi_{ap}\,\delta_{ap}}{\sum_{a=0}^{p} \pi_{ap}} = \sum_{a=0}^{p} \omega_{ap}\,\delta_{ap}, \qquad \omega_{ap} \equiv \frac{\pi_{ap}}{\sum_{a'=0}^{p} \pi_{a'p}}. \tag{9}$$

To make this mapping precise, we now define potential timing under IVF success and failure and link the age–horizon effects δ_{ap} and weights ω_{ap} to potential outcomes. Let $T_i(s)$ be the (calendar) year of first birth if success$_i = s$ (with $T_i(s) = \infty$ if no birth), and define the potential time-to-birth since IVF as

$$P_i(s) \equiv T_i(s) - IVF_i \in \{0, 1, 2, \ldots, \infty\}.$$

Appendix A.1 shows formally that (abstracting from conditioning), under $P_i(1) = 0$ and $\Pr(P_i(0) = 0) = 0$,

$$\delta_{ap} = \begin{cases} E\big[y_{ip}^{a} - y_{ip}^{\infty}\big], & a = p, \\ E\big[y_{ip}^{a} - y_{ip}^{\infty} \mid P_i(0) = p - a\big], & a < p, \end{cases}$$

where y_{ip}^{a} denotes the potential outcome at horizon p if the child's age at p were a (and y_{ip}^{∞} the no-birth potential outcome). For $a = p$, δ_{pp} is the ATT; for $a < p$ the δ_{ap} is the ATT for women who have their first born later on an unsuccessful IVF. The first stage coefficients equal

$$\pi_{ap} = \begin{cases} 1, & a = p, \\ -\Pr\big(P_i(0) = p - a\big), & a < p, \end{cases}$$

which shows that the weights for $a < p$ are negative.

The LPR-IV estimand at horizon p compares IVF successes and failures. As p grows, many women who failed the first IVF attempt nevertheless give birth later and enter the control group with younger children. The reduced form at horizon p therefore equals the contemporaneous age-p effect *minus* a first-stage–weighted average of earlier age effects. If earnings impacts are most negative when children are young, these negative weights mechanically offset the true effect and push $\hat{\gamma}_p$ toward zero as p increases. Interpreting the LPR-IV profile as genuine fade-out can therefore be misleading. In the next subsection we show how we can recover the dynamic, age-specific effects that the event study targets while exploiting exogenous variation from IVF success.

4.4 Estimating child-age effects (δ_a) with IV (Event-IV)

The classic event study identifies child-age effects δ_a by comparing outcomes of women who give birth at different times, assuming that timing variation is exogenous, and under a cohort-homogeneity condition. The mapping above shows that the LPR-IV estimand at horizon p is a first-stage–weighted average of age-specific effects for women who have different IVF-induced timing. We now show that a timing-homogeneity condition ($\delta_{ap} = \delta_a$ for $p \geq a$), also point identifies age-specific effects δ_a but now exploiting the exogenous variation of an IVF success.

To understand why timing-homogeneity is needed to recover the age-specific effects δ_a, note that at $p = 0$ women can only have a newborn ($a = 0$), so $\rho_0 = \pi_{00}\,\delta_{00}$, where δ_{00} is the average effect of having a zero-year-old immediately following IVF:

$$\delta_{00} = E[\delta_{i0}]$$

and under relevance ($\pi_{00} \neq 0$),

$$\delta_{00} = \frac{\rho_0}{\pi_{00}} \quad \text{is identified.}$$

At $p = 1$ a woman can have at most a one–year–old ($a \in \{0, 1\}$), so $\rho_1 = \pi_{01}\,\delta_{01} + \pi_{11}\,\delta_{11}$ where

$$\delta_{01} = E[\delta_{i0} \mid P_i(0) = 1]$$
$$\delta_{11} = E[\delta_{i1}]$$

Therefore, if we assume timing–homogeneity at age 0, $\delta_{01} = \delta_{00}$, i.e., the effect of having a zero-year-old is the same for complying women who give birth in $p = 0$ and $p = 1$, then

$$\delta_{11} = \frac{\rho_1 - \pi_{01}\delta_{00}}{\pi_{11}} \quad \text{is identified.}$$

By the same logic, we obtain a general recursive identification of δ_{pp} provided $\pi_{pp} \neq 0$, and under a timing–homogeneity (TH) condition that the effect at age a does not depend on IVF–induced delay:

$$\delta_{ap} = \delta_a \qquad \text{for all } p \geq a,$$

This assumption mirrors the cohort-homogeneity condition in the event-study design, and corresponds to the 'cohort-invariant dynamic effect' assumption in Ferman and Tecchio (2025) and Angrist et al. (2025).

We can estimate δ_a by imposing the timing-homogeneity assumption on equation (7) and and estimate it (and the corresponding first-stages (6)) using 2SLS in a panel where time p is indexed relative to the IVF attempt. However, to make the link to the event-study explicit we estimate the δ_a *in calendar time*. We can move from IVF-time (p) to calendar-time (t) by stacking horizons and instrumenting each child-age indicator with IVF success interacted with time-since-IVF dummies, while controlling for calendar time and time since IVF. This "Event-IV" mirrors the familiar event-study specification but relies on the exogeneity of IVF to supply the identifying variation, rather than exogeneity assumptions with respect to timing.

To see how we can change the centering from IVF-time p to calendar-time t, define the following time-since-IVF dummies $P_{it,p} = \mathbf{1}\{t - IVF_i = p\}$, and note that $y_{it} \equiv \sum_{p\geq 0} P_{it,p} y_{ip}$. Substituting (7) in this identity then gives

$$\begin{aligned} y_{it} &= \sum_{p\geq 0} P_{it,p}\Big[\sum_{a=0}^{p} \delta_{ap}\, \mathbf{1}\{a_{ip} = a\} + x_{ip}'\psi + \eta_{IVF_i} + \theta_p + u_{ip}\Big] \\ &= \sum_{a\geq 0} \delta_a \mathbf{1}\{a_{it} = a\} + x_{it}'\psi + \eta_{IVF_i} + \sum_{p\geq 0} P_{it,p}\theta_p + \varepsilon_{it}, \end{aligned} \tag{10, event-IV}$$

where the second line just re-centers from event time p to calendar time t: each observation (i,t) belongs to exactly one time-since-IVF cell $p = t - IVF_i$, so the child-age indicators remain unchanged. Imposing timing-homogeneity ($\delta_{ap} = \delta_a$ for $p \geq a$) collapses the stacked horizons to $\sum_{a\geq 0} \delta_a\, \mathbf{1}\{a_{it} = a\}$, while the IVF_i fixed effects and the flexible profile in time-since-IVF also stay unchanged.[10] The corresponding first stage follows in the same way. Multiply (6) by $P_{it,p}$ and sum over p to obtain the calendar-time first stage (for each age a):

$$\mathbf{1}\{a_{it} = a\} = \sum_{p\geq a} \pi_{ap}\,(\text{success}_i\, P_{it,p}) + x_{it}'\lambda_a + \zeta_{a,IVF_i} + \sum_{p\geq 0} P_{it,p}\,\theta_{ap} + u_{iat} \tag{11, FS, event-IV}$$

Adjusting for the time–since–IVF dummies $P_{it,p}$ ensures that each observation is compared only to observations at the same event time (the conditioning set for independence), and interacting success_i with $P_{it,p}$ creates one instrument per p, such that the first stages in (11, FS, event-IV) reproduce the *same* π_{ap} used above. Finally note that, with the full set of time–since–IVF dummies $P_{it,p}$ and because $t = IVF_i + p$, including IVF_i fixed effects is equivalent to including calendar–time

[10] Formally, with $P_{it,p} = \mathbf{1}\{t - IVF_i = p\}$, for each (i,t) there is a unique $p^* = t - IVF_i$ with $P_{it,p^*} = 1$ and $P_{it,p} = 0$ otherwise, hence $x_{ip^*} = x_{it}$ and $\mathbf{1}\{a_{ip^*} = a\} = \mathbf{1}\{a_{it} = a\}$. Timing-homogeneity implies $\sum_{p\geq 0} P_{it,p} \sum_{a\leq p} \delta_{ap}\, \mathbf{1}\{a_{ip} = a\} = \sum_{a\geq 0} \delta_a\, \mathbf{1}\{a_{it} = a\}$. Fixed effects pass through because $\sum_{p\geq 0} P_{it,p}\eta_{IVF_i} = \eta_{IVF_i}$ and $\sum_{p\geq 0} P_{it,p}\theta_p$ simply selects the relevant θ_p. Define the composite error $\varepsilon_{it} \equiv \sum_{p\geq 0} P_{it,p} u_{ip}$.

fixed effects τ_t; in practice we use τ_t (together with $P_{it,p}$) rather than IVF_i to align with the event-study specification.[11]

While the standard event study estimates effect relative to $a = -1$, equation (10, event-IV) uses the horizons $p \geq 0$ to estimate the child-age effects relative to the no–child state (i.e. $a < 0$ or $a = \infty$). This aligns the normalization with LPR-IV and makes the LPR-to-event mapping exact. This also means that IV pre-trends must be estimated separately in the pre–IVF period $p < 0$. In practice this makes very little difference because the exogeneity of success$_i$ balances potential outcomes in the pre–periods, and estimates relative to $a < 0$ closely match those normalized at $a = -1$.

It directly follows from the identification arguments above, that the *levels* of the relevant potential outcomes for the same complier groups are identified in the same way. We can therefore follow Abadie (2003), and estimate complier averages of y_{it}^{a} and y_{it}^{∞} by estimating (10, event-IV), which still includes the full set of mutually exclusive child-age dummies, with dependent variables $-\mathbf{1}\{a_{it} = a\}y_{it}$ (and $\mathbf{1}\{a_{it} \neq a\}y_{it}$). The coefficients on $\mathbf{1}\{a_{it} = a\}$ then recover estimates of y_{it}^{∞} (and y_{it}^{a}) for the same complier groups. This also allows us to scale the estimates of δ_a relative to the counterfactual y_{it}^{∞}.

Finally, the above suggests a simple check of the timing–homogeneity restrictions used in equation (10, event-IV). To identify δ_0 we use only $p = 0$ (no homogeneity needed). To identify δ_1 we add a single homogeneity assumption across the adjacent timing cohorts $p = 0$ and $p = 1$. Moving to δ_2 adds another two homogeneity assumptions: require that the δ_0 effect is invariant between $p = 0$ and $p = 2$ and that δ_1 is invariant between $p = 1$ and $p = 2$. In general, identifying $\{\delta_0, \ldots, \delta_P\}$ using information from horizons $p = 0, 1, \ldots, P$ requires $P(P+1)/2$ homogeneity assumptions. As P grows, these assumptions link complier groups that are progressively farther apart in IVF-induced timing. This delivers overidentifying restrictions: each δ_a can be estimated at multiple horizons $P \geq a$ and should be stable if timing–homogeneity holds. We implement this check by re-estimating the age profile for $P = 0, 1, \ldots, 11$. Figure A19 shows the age-specific effects are essentially unchanged as the horizon moves from 0 to 11, providing strong support for timing–homogeneity.

To summarize, *i)* centering time on birth renders the treatment invariant to dynamic extensive margin fertility responses over time, *ii)* adjusting for timing through $P_{it,p}$ accounts for the dynamic selection into the fertility attempt, and *iii)*

[11] We do not (and cannot) restrict the sample to eventual mothers, as doing so would condition on post–instrument outcomes and invalidate the instruments.

the instrumentation addresses potential remaining unobserved variable bias due to other sources of fertility.[12]

4.5 Definitions of fertility effects and the child penalty

We report all results as relative effects to ensure comparability with the literature (e.g. Kleven, 2022). Specifically, we scale the estimated fertility effects by the average counterfactual outcome that would have been observed at the same point in time in the absence of a child. For women, the estimand is

$$p_a^{women} \equiv \frac{E[y_{it}^a - y_{it}^\infty \mid a_{it} = a, \text{women}]}{E[y_{it}^\infty \mid a_{it} = a, \text{women}]} = \frac{\delta_a^{women}}{y_a^{\infty,women}},$$

where $y_a^{\infty,women}$ denotes the average counterfactual outcome for mothers of child age a. In the event-study specification, this counterfactual is obtained by subtracting the estimated fertility effects δ_a from the observed outcomes of parents with children of age a. In the IV specification, we follow Abadie (2003) to recover the corresponding counterfactual outcomes. We construct the same measures analogously for partners.

We define the (scaled) child penalty as the difference in relative fertility effects between women and partners:

$$p_a = \frac{\delta_a^{women}}{y_a^{\infty,women}} - \frac{\delta_a^{partner}}{y_a^{\infty,partner}}. \tag{12}$$

Standard errors for the rescaled effects are obtained using the Delta method (see Appendix A.2).

5 Instrument validity

For the instrumental variable – success in IVF treatment – to be valid, it has to be uncorrelated with any determinant of the outcomes we study. The testable implications of this assumption are investigated in Table 2. Here, we report estimates from a regression of pre-IVF earnings (column 1), and IVF success (column 2), on a number of observable predetermined characteristics capturing women's

[12] Note that, while we instrument having a child of a particular age at different times since IVF, we only use one randomization. Challenges related to multiple instruments are therefore not relevant here. Nonetheless, we have also estimated the model using multiple IVF attempts as separate instruments, and the results remain virtually unchanged. Additionally, we have estimated the model by IVF attempt, again finding virtually identical estimates. This supports the homogeneity assumption. Results are available upon request.

Table 2. Instrument validity

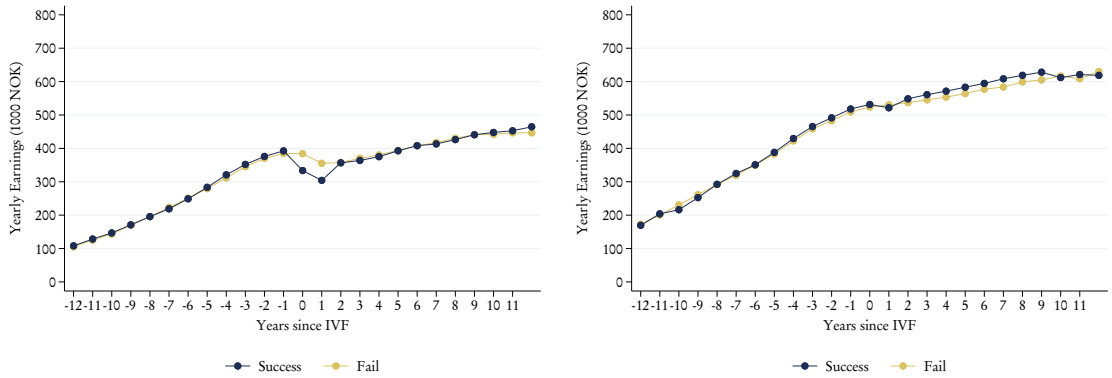
	Pre-IVF Earnings (100K NOK) (1)		IVF Success (2)	
	est.	*s.e.*	*est.*	*s.e.*
Woman characteristics				
Earnings (100K)			0.004	(0.003)
Hours (FTE)			-0.005	(0.010)
Sickness absence days (/10)	-0.010	(0.002)	0.001	(0.001)
GP visits	-0.036	(0.005)	-0.002	(0.002)
Psychological symptoms	-0.126	(0.034)	0.006	(0.010)
Hospital days (/10)	0.000	(0.002)	-0.001	(0.001)
Partner characteristics				
Age (/10)	-0.221	(0.031)	-0.003	(0.010)
Earnings (100K)	0.115	(0.008)	0.001	(0.002)
Hours (FTE)	-0.273	(0.045)	-0.008	(0.012)
Education (ref. master)				
- Compulsory	-0.134	(0.057)	-0.021	(0.018)
- High School	-0.147	(0.053)	-0.030	(0.015)
- Bachelor	-0.051	(0.054)	0.001	(0.015)
Constant	3.487	(0.247)	0.361	(0.094)
Mean dependent variable	3.38		0.31	
Joint F [p-value]	44.3 [<.001]		1.3 [0.228]	
N Women	10 033		10 033	

Note: This table reports estimates and standard errors from a regression of pre-IVF earnings (column 1), and of IVF success (column 2) on a number of observable predetermined characteristics capturing women's demographics, labor market attachment and health. Labor market and health variables are averaged over the four years preceding the first IVF trial. Education is measured in the calendar year before the IVF attempt, and age is measured at the date of the attempt. Missing variables are set to 0, and in these cases we include a dummy equal to 1 if replaced, zero otherwise. As in the event-IV specification in equation (10, event-IV) and (11, FS, event-IV), both regressions include dummies for calendar time, time relative to IVF treatment, woman's age, and education. Joint Fs [p-value] refer to tests of joint significance of the characteristics shown in the table.

demographics, labor market attachment and health.[13] As in the 2SLS specification in Equation (10, event-IV) and (11, FS, event-IV), all regressions include controls for calendar time, time since IVF treatment, maternal age, and education, which are known predictors of success (CDC, 2012; Groes et al., 2017). Our results therefore rely on conditional exogeneity of success, and not on an assumption that success is unconditionally random.

In column (1), the regression of pre-IVF earnings on background characteristics

[13] In appendix Table A1 we also show the raw means by success at first trial.



(a) Women **(b)** Partners

Figure 2. Average Earnings Relative to IVF Attempt, by IVF Success

Note: Estimates are regression-adjusted for calendar year, maternal age, and education. Predicted earnings by IVF success and time are averaged over the covariate distribution in the estimation sample. The sample includes all women (and their partners) who had their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (unique women = 10,033; observations = 173,480).

highlights potential confounders of our instrument. Many of these characteristics are strongly correlated with earnings (our main labor supply measure): women with poorer health, as measured by visits to their primary care physicians for any reason or for psychological symptoms, have lower earnings. Women whose partner has higher earnings also have higher earnings themselves. All characteristics are jointly significant in explaining pre-earnings, with a joint p-value that is smaller than 0.001.

A necessary condition for conditional exogeneity is that IVF success is not correlated with all observable characteristics that predict earnings. Column (2) indicates that characteristics predictive of earnings pre-randomization are generally not predictive of the instrument. For example, while hospital days is marginally associated with the IVF success rate, it is not predictive of earnings. Moreover, a test for joint significance of all variables is not significant and renders a p-value of 0.23.

These results are consistent with the instrument, success, satisfying exogeneity conditional on women's age and education. Any remaining confounder must be correlated with potential earnings and uncorrelated with pre-earnings up to twelve year prior to the IVF attempt. While it is theoretically impossible to rule out the existence of important potential confounders, we struggled to come up with concrete examples.

Although Table 2 indicates that any imbalance is likely to be minor, this test is based on an average over the four years preceding the first IVF trial. To make

sure that this average does not hide any imbalance in *trends*, Figure 2 plots average earnings for each year since the first IVF trial, by success, completely adjusting for the controls included in our main specification: calendar time, maternal age, and maternal education. More precisely, we first construct the estimates in the figure stratified by calendar year, maternal education and maternal age. We then compute the population level estimates by averaging across cells for each year since the first IVF trial. We see that to the extent that there is an imbalance it is constant over time and trends in earnings are essentially identical in the 12-year period leading up to the trial. We interpret these results as lending strong support for the assumption that the results of the IVF is indeed conditionally as good as random.

For the exclusion restriction to hold, we require that IVF success affects no variables other than fertility directly. Some argue that disappointment and related outcomes constitute a violation of exclusion (e.g. Gallen et al., 2023). While women who fail to conceive following IVF may experience disappointment, depression, or divorce (e.g. Bögl et al., 2024; Martinenghi and Naghsh-Nejad, 2025), we argue that such outcomes operate through fertility rather than as a direct effect of the trial and therefore do not violate exclusion. We discuss this in detail, along with issues of external validity, in Section 8.

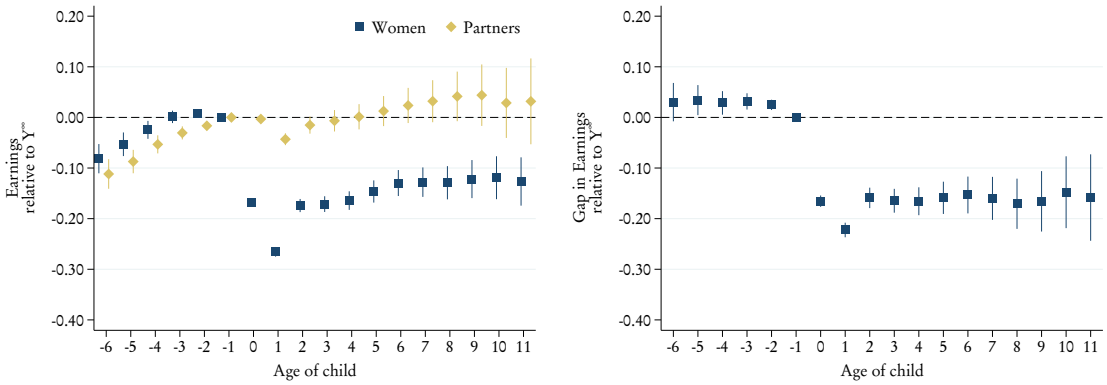
6 Children and labor market outcomes

We now present the estimated effects on earnings for the three different models described in Section 4: the standard event-study, the instrumental variable effect estimates of fertility since the IVF attempt (LPR-IV), and our specification that combines these two approaches (event-IV). For each model, we report scaled estimates for women, partners, and the gap between the two (women minus partners, as in equation 12). Our discussion of long-term effects will focus on age 6, as our panel is balanced up to and including this age. However, we also present effects up to age 11, though most of these results are confined to the appendix.

6.1 Event-study estimates

We start by reporting the results using the regular event-study specification of equation (2), estimated on IVF-women and their partners in Figure 3(a).[14] Both women and partners display a comparable pre-trend leading up to birth, indicating that those who have children earlier are on relatively steeper age-earnings profiles

[14]This means we include only IVF women who eventually have children, following standard practice in the event study literature. The non-IVF sample already consists of mothers only.



(a) Women and partners **(b)** Gap

Figure 3. Earnings – Event study estimates

Note: OLS event study estimates from specification (2). Panel (a) shows effects separately for women and partners, panel (b) shows the difference between women and partners. Estimates are scaled relative to each gender's counterfactual earnings (Y^{∞}), as described in Section 4.5. Point estimates are presented in table form in Table A4. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were registered with a partner at the time, and eventually had at least one child (observations = 145,571; unique women = 8,349).

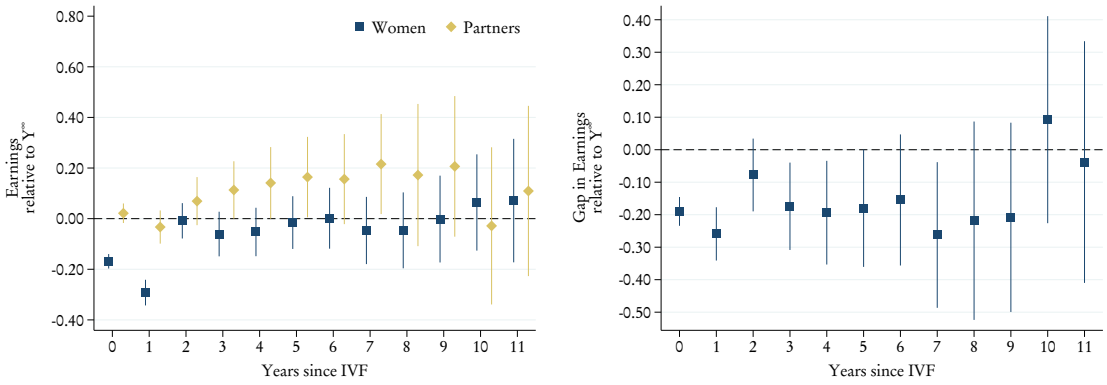
compared to those who have children later. Following birth, IVF women see a sharp drop in earnings of about 27 percent which then attenuates somewhat and stabilizes at around 13 percent in the longer run. Partners, in contrast, experience almost no negative effects on earnings following childbirth. Rather, they see a small increase of about 2 percent in the longer-run.

Figure 3(b) shows the corresponding effects on the earnings gap between women and their partner. As both parents follow a similar upward-sloping trend in earnings there is no discernible pre-trend, but there is still a substantial difference after birth at around 15 percent, which in the long-run is almost entirely driven by the drop in women's earnings.

We repeat our analysis on a sample of non-IVF women in Section 8, and find similar results. The estimates for IVF women are therefore in line with existing event-study evidence from Norway (Andresen and Havnes, 2019; Andresen and Nix, 2022) and comparable countries such as Denmark (Kleven et al., 2019).

6.2 LPR-IV estimates

We now turn to the estimated earnings effects of fertility using the LPR-IV model described in equation (4) with the outcome of the IVF treatment as the instrumental variable. Appendix Figure A1 reports the estimated first stages from equation (4),



(a) Women and partners **(b)** Gap

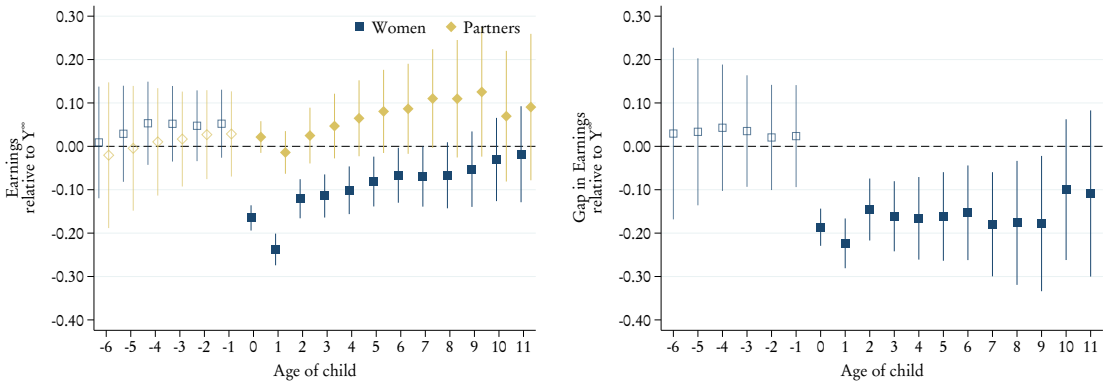
Figure 4. Earnings – LPR-IV estimates

Note: Estimated effects of fertility on earnings using the LPR-IV model described in equation (4) on our data. Panel (a) shows effects separately for women and partners, panel (b) shows the difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in Section 4.5. Full set of point estimates are reported in Table A5 in the appendix. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033).

essentially the difference between the average fertility rates between successful and failed IVF attempts shown in Figure 1. By construction, the first stage equals 1 nine months after the IVF treatment. It then declines over time as always-takers realize fertility. By the end of the first year, the first stage coefficient is already below 0.8, before stabilizing at 0.2 in the longer run. Despite this decline, the estimates are all highly statistically significant: Women who are successful in their first IVF-trial are therefore always more likely to have children than those who failed their first trial. The F-statistic is never below 500 in the first nine years since IVF and are reported in appendix Table A10.

Figure 4(a) shows the IV estimates of equation (5), separately for women and their partners. Women's earnings drop by about 30 percent in the year following the IVF treatment, but the effect quickly reverts to zero in the second year, at which level it remains for the remaining period. For comparison, Lundborg et al. (2017) find long run earnings losses for women at around 11 percent. As discussed in Section 4, this estimate is probably biased toward zero (i.e. the actual effect is more negative than the estimate) since delayed fertility is confounding the counterfactual earnings profile and introduces a positive bias.

Partners see no earnings drop immediately following IVF treatment. If anything, there is an earnings premium of about 16 percent six years after birth. While the estimates are increasingly noisy they appear to be stable, estimating the average



(a) Women and partners **(b)** Gap

Figure 5. Earnings – Event-IV estimates

Note: Estimated effects of age of child on earnings using the event-IV model described in equation (10, event-IV). Panel (a) shows effects separately for women and partners, panel (b) shows the difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 4.5. Full set of estimates in table form are reported in Table A6 in the appendix. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033).

impact for a three year window around year six gives an estimate that is significant at conventional levels. Figure 4(b) reports the estimated effect of fertility on the earnings gap between women and their partners. This fluctuates a bit over time, averaging at 15 percent after six years, driven exclusively by the positive point estimate for partners' earnings.

While the event-study and LPR-IV models produce similar estimates of the fertility effect on the earnings gap between women and their partners, the individual point estimates for women and for partners are strikingly different. Where the event-study finds that women's earnings fall in the neighborhood of 13 percent, the LPR-IV specification shows that earnings reductions are substantial only on the very short run and essentially zero after two to three years. Direct comparison of these estimates is however complicated because they do not recover the same effects. We therefore now turn to our IV event-study results which reconcile these approaches.

6.3 Event-IV estimates

Figure 5 presents the estimated effects of children from our event-IV specification as described in equation (**??**). F-statistics for the first stages are reported in appendix Table A10 and far exceed conventional levels for statistical significance. In Figure 5(a) we see that while we estimate an immediate drop in women's earnings of about

24 percent, the long run earnings loss is around 7 percent. This is half of the effect on earnings estimated in the event-study model and the differences are statistically different at conventional significance levels. We also see no signs of any anticipation effects in the years leading up to the trial. No meaningful earnings drop is seen for partners around childbirth. In contrary, the estimates suggest an increase in earnings over time, reaching around 9 percent in the long run. Figure 5(b) plots the estimated gap between women and partners from the event-IV model. There is no evidence of an earnings gap before birth, at which point it drops to around 20 percent, before stabilizing at around 15 percent in the longer run. This long run parental earnings gap is primarily driven by the partners.

For completeness, appendix Figures A6 and A7 report results for additional labor market outcomes (hours, employment and hourly wages) for the event-IV model.[15] The broad takeaway from the event-IV estimates is that for women the results on the long run seem to be mostly driven by responses at the employment margin. While disentangling intensive and extensive margin responses is not possible without a structural model, the results in Figure A7, which condition on employment, suggest that while women reduce hours on the short-run, their hours responses on the longer run appear to be negligible. Similar results for hourly earnings also give little sign that there are sizeable long-run effects. For partners, we also find employment responses. Contrary to women, the results in Figure A7 suggest that there are some long-run impacts on hours and hourly earnings, where the latter may be explained by career returns.

Finally, there are several major welfare programs in Norway that aim to replace lost labor market earnings through provisions such as parental and sick leave benefits. Our main earnings measure does not capture these welfare benefits, nor does it cover earnings for self-employed persons. We therefore supplement our main findings using an extended income definition that includes these sources. Appendix Figure A11 shows that this, as expected, dampens the estimates in the very short run, but does not affect our longer-run estimates.

7 Reconciling estimates of the effect of fertility

Table 3 summarizes the estimates for the three models by reporting the long-run estimates of earnings for the woman, the partner, and the gap between the two known as the child penalty. Long–run estimates are evaluated when the child is six years old ($a = 6$) which is the last age for which we have a balanced panel. We

[15] Results for the same outcomes for the other models are reported in Figures A2 - A5.

Table 3. Comparison of long-run (age 6) fertility effects and child penalty estimates across models

	LPR-IV (1)	Event (2)	Event-IV (3)	Event vs. Event-IV (2) - (3)
Gap	-0.15	-0.15	-0.15	-0.00
	(0.09)	(0.02)	(0.06)	(0.05)
Woman	0.00	-0.13	-0.07	-0.06
	(0.06)	(0.01)	(0.03)	(0.03)
Partner	0.16	0.02	0.09	-0.03
	(0.09)	(0.02)	(0.05)	(0.05)

Note: Table shows estimates of earnings for woman, partner, and the gap (woman - partner), evaluated at $a = 6$ ($p = 6$ for LPR-IV). Column (1) shows estimates from the LPR-IV model, column (2) shows the estimates from the event-model, column (3) shows estimates from the event-IV model, and column (2) - (3) shows the difference between the event model and the event-IV model. Standard errors for gaps between parents and differences across models are bootstrapped using 199 repetitions. The sample for the IV estimates includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033). The sample for the event study estimates includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were registered with a partner at the time, and eventually had at least one child (observations = 145,571; unique women = 8,349).

report analogous results for age eleven ($a = 11$) in appendix Table A7.[16] Column (1) shows estimates from the LPR-IV model, column (2) shows estimates from the event model, column (3) shows estimates from the event-IV model. The final column compares the estimates from the event model with the event-IV model. This difference can be interpreted as the bias present in the event estimates under the assumptions of the event-IV model and absent notable complier heterogeneity which we document below.

The first thing to note is that the estimates of the impacts of fertility on the earnings gap between women and partners are sizable in the three different models. All three models estimate a long-run impact on the parental earnings gap of 15 percent. But where the LPR-IV model suggests that none of this gap is driven by women, the standard event study, in contrast, finds large negative and statistically significant effects on maternal earnings and a small positive estimate for partners.

The estimate for the long-run parental earnings gap from the event-IV specification is identical to that of the LPR-IV model. However, when it comes to the separate estimates for women and partners, the event-IV model paints a different picture than the event-study model. For women, it estimates only a small long-run

[16]Appendix Tables A8 and A9 show that removing the restriction to partnered women in the year before IVF gives estimates nearly identical to the baseline.

negative impact of children on earnings of 7 percent, compared to 13 percent in the event-study model. For partners, the event-IV model estimates an earnings increase of around 9 percent compared to 2 percent in the event-study model.

The estimates for age 11 in Table A7, paint a very similar picture. The estimated gaps in the event models appear to be stable, and the event-IV estimates suggest that the effect for women are smaller, while the partner effects remain positive or are even increased. Note however that our sample is no longer balanced and is much reduced. The estimates are consequently very noisy which means that we cannot reject that the effects are the same as for age 6.

These results illustrate that the estimates, interpretation and policy implications of the fertility effects not only depend on whether one considers the gap between parents or the impact on women or partners separately, but also on which particular model is applied. This raises the question of what drives these differences, and we therefore now delve deeper into the underlying causes.

7.1 Event-IV and LPR-IV

We report the estimated weights in (9) for $p = 6$ in Figure 6. On the left-hand y-axis we plot the first-stage coefficients for having a child of age a six years after IVF (π_{a6}). The right-hand y-axis shows the normalized weight for each first-stage (ω_{a6}). The figure shows that there is a large positive weight for $a = 6$ which means that when estimating the fertility effect on earnings, the LPR-IV estimator puts a large positive weight on the effect of having a child p years old. However, the estimated effects for having a child any younger than six years old (i.e. $a < p$) are given a negative weight. As the effect of having children is negative, this weighting biases the fertility estimates in the LPR-IV model towards zero relative to the contemporaneous effect of having a child within a year from the IVF attempt, which has a positive weight. We show that this pattern holds for all p in appendix Figure A12 and A13. On the very short run ($p = 0$) the fertility effect γ_0 is equal to the earnings effect δ_0, but with time the contemporaneous earnings effect δ_p gets an increasingly smaller relative weight.

We can use our event-IV estimates of δ_a and π_{ap} to construct alternative estimates of γ_p and compare these to the estimates of γ_p based on the LPR-IV estimates from equations (4) and (5). The mapping is illustrated in appendix Figure A14 where we plot the results for women's earnings from the LPR-IV model along with the estimates constructed from the reduced form and the first stages from our event-IV. Reassuringly, these results confirm the equivalence between the reduced forms, confirming that the results are indeed only differing due to our decomposition of

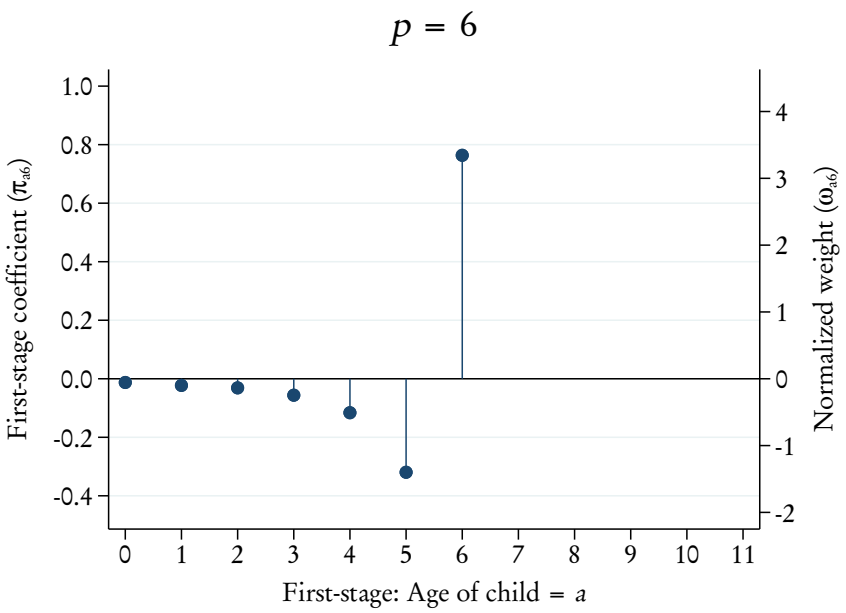


Figure 6. Mapping the first stages of LPR-IV to event-IV

Note: This figure shows how the first stage coefficient in the LPR-IV model six years after the IVF trial can be defined as a weighted average of the first stages for having a child of 6 years or younger in the event-IV model. The left y-axis plots π_{a6} (the first stage coefficients by age a for potential age $p = 6$), while the right y-axis shows $\omega_{a6} \equiv \pi_{a6} / \sum_{p \geq 0} \pi_{a6}$ (the normalized first stage coefficients by age a for potential age $p = 6$). The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033).

fertility into dynamic treatment effects of having a child of a specific age.

7.2 Event-IV and Event

The estimates for the earnings effects differ across the event-study model and our event-IV model. We now investigate the sources of these differences. We focus on how a violation of the exogeneity assumption in event study models leads to overestimated effects of fertility on earnings for women (and their partners).

The validity of the estimates produced by the event-study model shown in Figure 3 depends on the assumption that women do not time fertility to their unobserved counterfactual earnings trajectory conditional on observed age and time. Ideally, one would like to compare prospective mothers with women who have similar intended fertility timings but where the subsequent birth is as good as exogenous. Our IVF data provide us with such timing information since we know the date at which women insert their fertilized egg. In Figure 7 we show how the standard event study estimates are affected by adding dummies for time since the first IVF trial to the standard event model of equation (2).

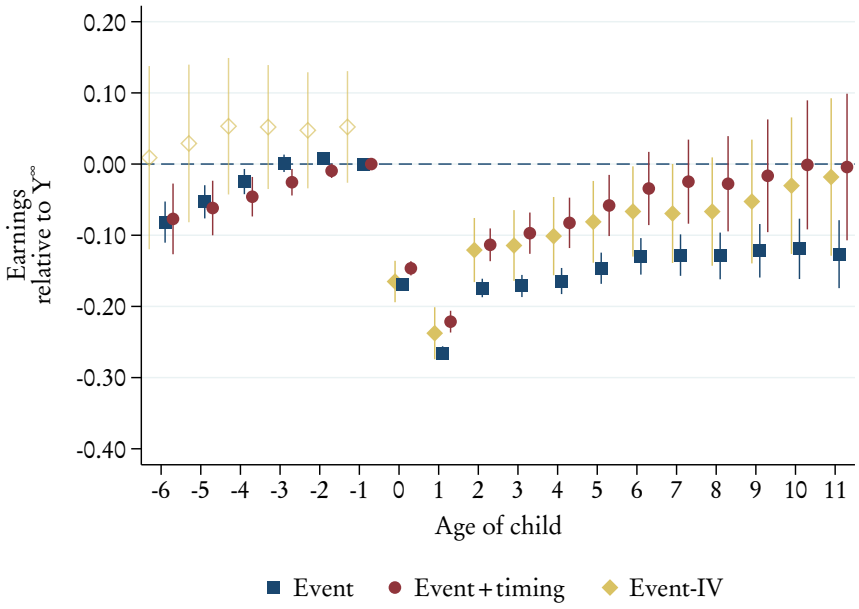


Figure 7. Event vs. Event-IV estimates

Note: This figure compares estimates from the event-study specification with and without controls for time since IVF trial, to results from the event-IV specification. All estimates are scaled relative to counterfactual earnings (Y^{∞}) as described in section 4.5. The IV sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033). The event-study sample is the subset who eventually had at least one child (observations = 145,571; unique women = 8,349).

As seen in Figure 7, controlling for timing has little impact on pre-trends. Meanwhile, there is a significant reduction in post-birth effects of having a child on earnings. Where the earnings reduction for women was about 13 percent in the standard event-study setup, controlling for timing almost eliminates the penalty to about 3 percent after 6 years, and completely by year 11.

To provide more insight on how adjusting for timing affects the results, Figure 8 reports estimated counterfactual earnings normalized to $\tau = -1$ for the event-study with and without controlling for timing. Y^a is the predicted earnings profile in the presence of a child of age a, while Y^{∞} is the predicted earnings profile in absence of a child. The estimates of Y^a and Y^{∞} from the event-study model without timing show that women face on average upward sloping earnings until their pregnancy, followed by a sharp drop in the first year after birth. Earnings growth then recovers and after three years mothers appear to be back on a new age-earnings profile on a lower level, but comparable slope, such that there is a permanent and constant wedge between wages for women with and without children.

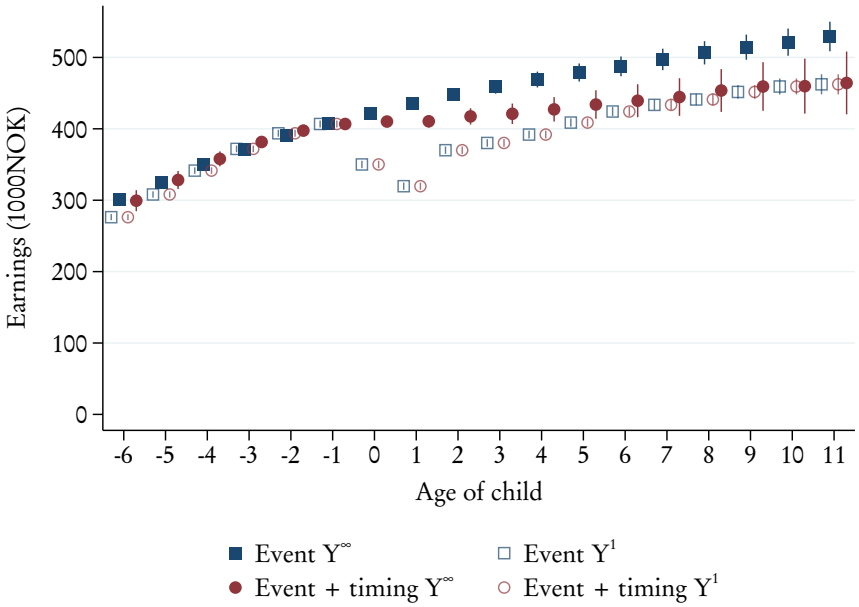


Figure 8. Counterfactual earnings profiles – Event-study estimates.

Note: This figure shows the estimated potential earnings without child (Y^{∞}), and with child (Y^{a}), as estimated from the event-study model, with and without controls for time relative to first IVF attempt. The event-study sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were registered with a partner at the time, and eventually had at least one child (observations = 145,571; unique women = 8,349).

The counterfactual earnings profile without a child, Y^{∞}, is marginally flatter leading up to (counterfactual) birth and continues to grow beyond that time. The difference between this earnings profile and Y^{a} is the estimate for maternal earnings in the standard event-study specification. These estimates rely however on a comparison of women with different intended fertility timing. After taking these ex-ante differences into account in the estimation of "Y^{∞} + timing" the earnings profiles are now nearly aligned leading up to birth. Crucial for the estimates, women appear to have children when the growth rate of counterfactual earnings (Y^{∞} + timing) starts to decline, and their earnings are therefore ultimately lower than those of women who have children later. The standard event-study specification does not capture these differences and consequently overstates the estimated effects on maternal earnings and the earnings gap.

In a final step, we compare the event study estimates that control for timing to the full event-IV estimates. Figure 7 shows that once we control for timing in the event-study model, the fertility effect estimates are much more similar to our event-IV model estimates – to the extent that the differences are no longer statistically significant. This is not surprising: there are by construction no never-takers to

our instrument, and had there also been no always takers, that is, if women could not have children without an IVF treatment, then the event-study and event-IV estimates are identical after controlling for the endogenous component of fertility, namely the timing of the fertility attempt. In our application, as much as 80 percent of fertility is channeled through the IVF treatment, which means that the compliers to our instrument are very similar to the population that provides the identifying variation in the event model that adjusts for the timing of the fertility attempt. This is also shown in appendix Table A11 which reports population and complier statistics using Abadie κ-weighting (Abadie, 2003). Compliers are almost identical to the full sample across all characteristics. These findings suggest that although our event-IV estimates technically are local average treatment effects they are likely very similar to the average treatment effect in the presence of treatment-effect heterogeneity. These results also imply that even though we use the event-IV estimates as a benchmark, none of our main findings crucially depend on the instrumental variable assumptions.

7.3 Alternative event-study estimators

Event studies often assess the credibility of the exogeneity or parallel-trend assumption by evaluating the pretrends. Rambachan and Roth (2023), for example, formalize a robustness check based on the idea that pre-trends are informative about violations of parallel trends and propose checks to assess how sensitive results are to deviations from the pre-trends after treatment. Appendix Figure A17 reports event-study estimates that adjust for the baseline of a linear extrapolation of the pre-trend into the post period. The figure shows that the adjusted results exacerbate the bias relative to the standard event-study specification. The reason is that the sign of the selection bias reverses after birth as seen in Figure 8, which results in counterfactual earnings estimates that are even higher with extrapolated pre-trends than in the standard event-study.

In the traditional event-study model, both previously treated and untreated observations are used to estimate the counterfactual for a treated unit at any point in time. This is a valid approach only under the assumptions implicit in the model specification of equation (2), namely treatment effect homogeneity and the correct specification of the counterfactual earnings profiles defined by the model. Recent advances in econometrics have shown that violations of these assumptions in conventional event-study estimators can severely bias effect estimates (Borusyak et al., 2024; Goodman-Bacon, 2021; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; de Chaisemartin and D'Haultfœuille, 2020). In the context of the

impact of children on earnings, the treatment effect homogeneity assumption is violated if children have a larger effect on parental earnings when they are younger, as suggested by Figure 3. An additional violation occurs if there is a selection on gains in the timing of fertility, for example if women time their fertility based on the effects on earnings.

To assess whether more flexible event study estimators that account for heterogeneity in treatment effects recover earnings estimates that are consistent with our event-IV model (or the event-study model that adjusts for time-since-IVF), we apply the estimator Callaway and Sant'Anna (2021) to our sample of IVF women. This estimator constructs all possible two-by-two treatment cohort specific contrasts relative to the last pre-treatment period and aggregates these up to average-treatment effects on the treated (see Roth et al., 2023, for a review).[17] We allow for heterogeneity by women's age at birth (rather than calendar year).

Results are plotted in appendix Figures A15, along with the conventional event study estimates. Figure A15 shows that for both women and their partners the Callaway and Sant'Anna (2021) approach lines up the pre-trends, but estimates larger negative effects of children on earnings. This is consistent with the results based on the extrapolation of pre-trends above.

These findings highlight that the type of selection that we documented using the event-study specification which allows counterfactual earnings profiles to depend on fertility timing (Figure 7) is hard to capture without having information about the timing of fertility intent, and provide a cautionary tale on what can be learned from the presence (or absence) of pre-trends.

8 External validity

Our results are based on IVF women, who make up around six percent of all births in Norway. This is a sizable and growing group, and therefore of interest in its own right. The advantage of this setting is that it allows us to exploit two key features uniquely provided by the IVF context: i) detailed information about the timing of fertility and ii) conditional randomization of birth. However, an important question is whether our findings also speak to the population at large. Generalization rests on two questions: (i) whether IVF women differ from other women, and (ii) whether IVF births differ from other births. We discuss these in turn.

[17] We use Stata's implementation of Callaway and Sant'Anna (2021).

8.1 Are IVF women different from regular women?

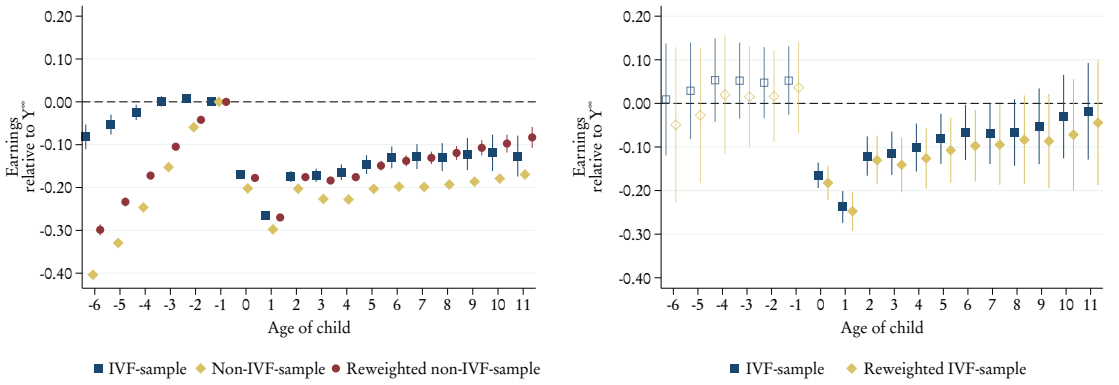
A key question for external validity is whether IVF women differ from non-IVF women in ways that change the labor-market response to children. Two channels matter: selection–if the timing of first birth is related differently to potential earnings profiles–and treatment heterogeneity–if the causal effect of having a child differs between the groups.

We start by comparing fertility timing across samples. Table 1 shows that women undergoing IVF are older and more educated at conception and, on average, have slightly fewer children. These intensive-margin differences are fully accounted for by age and education. When we reweight the non-IVF sample to match the IVF distribution of age and education at conception, completed fertility among those with at least one child is virtually identical across groups (Table A2).[18] Table A3 reports additional characteristics for the reweighted non-IVF sample. Conditional on age and education, IVF women (and their partners) have somewhat higher earnings, and IVF women exhibit slightly higher sickness absence. IVF couples therefore appear modestly positively selected, but the remaining differences are small.

An alternative way to assess population heterogeneity is to compare standard event-study estimates for IVF women with those for non-IVF women, and to diagnose differential selection or anticipation from pre-trends. Figure 9a reports estimates for both groups. Pre-trends are noticeably steeper in the non-IVF population, consistent with stronger selection. This evidence aligns with the Norwegian Mother and Child Cohort Study (Magnus et al., 2006), where 82 percent of mothers report a planned pregnancy, which implies that roughly one in five births are unplanned in the general population. Earlier unplanned pregnancies are plausibly associated with more negative selection, in line with Gallen et al. (2023) who, using Swedish data, find larger negative effects for younger women. In contrast, IVF conceptions are actively planned and timing is more constrained by treatment schedules, leaving less scope for selective timing. Taken together, these patterns indicate that timing bias is likely a concern in the broader population as well.

A comparison of the post-birth fertility effects can also shed light on heterogeneity across populations. Figure 9a shows that the non-IVF women experience a larger drop in earnings following birth than the IVF women. However, reweighting the non-IVF sample makes the event-study estimates of responses to children remarkably similar to those of the IVF-sample. This exercise suggests that there is little evidence that IVF women's response to childbearing (or bias) is very different from women in

[18] Observations are weighted by the inverse propensity score, with scores estimated via a probit fully saturated in age and education indicators.



(a) Event-study specification **(b)** Event-IV specification

Figure 9. Earnings. Reweighted estimates

Note: Panel (a) shows event-study estimates for the IVF sample (145,571 observations; 8,349 women), the non-IVF sample (1,972,754 observations; 109,791 women), and the non-IVF sample reweighted to match the composition of the IVF group. All samples include women who eventually have at least one child. Panel (b) shows event-IV estimates for the IVF sample and for the same sample reweighted to match the composition of non-IVF women. The event-IV sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children beforehand, and were registered with a partner at the time (173,480 observations; 10,033 women). Weights are inverse propensity scores estimated from a fully saturated probit model in age and education. Estimates are scaled relative to each gender's counterfactual earnings (Y^{∞}), as described in Section 4.5.

the population at large who are comparable in age and education.

To examine whether causal effects are heterogeneous by age and education, we additionally reweight the IVF sample to match the composition of non-IVF women before re-estimating the event-IV model. Since selection bias is eliminated in these specifications, any observed differences should reflect effect treatment effect heterogeneity alone. Figure 9b shows that this adjustment only marginally changes the estimates for impacts of children. This implies that while impacts are somewhat larger for less educated and/or younger women, effect heterogeneity across these characteristics would not lead to dramatically different estimates for the overall population.

We interpret these results as indicating that observable heterogeneity is unlikely to explain the gap between the event-IV estimates for IVF women and the standard event-study estimates for non-IVF women. Given that event-study profiles are similar across groups and that removing selection bias attenuates women's earnings losses in the IVF sample, it is likely that event-study estimates for non-IVF women are also overstated (too negative). However, without a valid instrument for the non-IVF population, we cannot quantify the magnitude of this bias.

8.2 Are IVF births (and non-births) different from regular births (and non-births)?

A frequently noted phenomenon in the context of IVF is the presence of a "disappointment effect" following an unsuccessful IVF, which could affect behavior or mental health (e.g. Gallen et al., 2023; Martinenghi and Naghsh-Nejad, 2025). In our data, we see short-lived changes around treatment, concentrated among a minority of women. In particular, the probability of having at least one mental-health–related medical visit per year is about 12 percentage points lower for those who become mothers compared to those who do not (Figure A8a). The impact on *severe* psychological symptoms is modest: the probability of having at least one such visit declines by about 2 percentage points among mothers relative to non-mothers, an effect that is significant only in the first year after childbirth.

Impacts on other non–labor-market outcomes, such as marital stability, are also modest in the data. Figure A8b shows that having children slightly reduces divorce risk over time, and this has a negligible influence on our earnings estimates (Figure A9).

The mental-health and marital-stability responses are operating through the treatment–having a child–rather than through the instrument itself, and therefore do not threaten the exclusion restriction for the post-birth effects we estimate, but are rather direct consequences of childbearing. Separating these mediating outcomes from the other ways in which having a child affects labor market outcomes requires additional instruments which we do not have. However, estimates of the effects of children are virtually unchanged when we control for these channels (Figure A9) or exclude potentially affected women (Figure A10), suggesting that they do not play a major role here.

Disappointment from not conceiving is unlikely to be specific to IVF. Standard event-study designs in the broader population will also include unsuccessful conception attempts, and thus disappointment, in the counterfactual group.[19] This is especially true for our event-study estimates for IVF women, where the estimation of the counterfactual disproportionately relies on failed IVF attempts. Disappointment is therefore unlikely to explain the differences between the event-study and the event-IV results. Finally, the near invariance of the age-of-child estimates as we expand the estimation horizon (Figure A19) is difficult to reconcile with any large and persistent disappointment-driven effects, as these would typically induce systematic drift across horizons, which we do not observe.

[19] Comparable issues arise in other IV settings where eligibility rules or lotteries may generate disappointment among non-recipients, such as school assignment, housing assistance, job training, or health-care access.

To summarize, while IVF births clearly differ from natural births in some dimensions–and in particular their medicalized context–we interpret these facts as suggesting that disappointment-related mental-health responses are unlikely to play a major role in driving our main results, and that there is little evidence that births following IVF are fundamentally different from regular births, especially in the long run. While we therefore expect the same qualitative mechanisms and potential biases to operate more broadly, even though their quantitative importance in other populations remains an open question.

9 Conclusion

Social scientists and policy makers have devoted considerable effort to understanding the drivers of the gender wage gap. In particular, significant attention has been paid to how parenthood, specifically motherhood, can be a key driver of this disparity. A broad conclusion coming of this work is that women experience an abrupt and permanent drop in earnings after becoming mothers, whereas their partners' earnings remain largely unchanged. The resulting increase in the earnings discrepancy between mothers and fathers following parenthood is commonly referred to as the child penalty.

Empirically, much of the heavy lifting in this literature is done by the event-study framework. The current paper contributes by assessing the validity of the key assumptions in the event-study specification commonly used for identification. We exploit external identifying variation coming from information on the timing and randomness in the success rates of IVF treatments.

Standard event studies compare women who have children to women of similar age who have children later in life. Using data on Norwegian women undergoing such treatments, we find that women time fertility as their earnings profile flattens. The implication of this is that the event-study overestimates mother's earnings penalty as it relies on estimates of counterfactual wage profiles that are too high. Accounting for the timing of the fertility attempt in the event study substantially reduces the earnings effects of fertility. Using success at IVF trials to instrument for fertility takes any remaining endogenous sources of fertility into account, but this does not substantially change our conclusions. We estimate longer-run earnings effects for women of around 7 percent, which is half of the effect size uncovered by a standard event-study setup in the same sample. Even though we use the event-IV estimates as a benchmark, these results also imply that none of our main findings crucially depend on the instrumental variable assumptions of this model.

Our approach builds on the setup of Lundborg et al. (2017) who also use an IV strategy for women undergoing IVF treatments. Using their specification we find large positive point estimates for partners and no evidence of effects on women in the longer-run. We show that relative to the event-IV approach centered on birth, their IVF-attempt-centered estimator provides estimates that are mixtures of the effects of having children of various ages where, with time, the model puts increasing negative weight on the effect of children born after the first IVF trial. We therefore decompose the estimates of Lundborg et al. (2017) into plausibly causal analogues of the parameters targeted by the event-study model.

While the effects on the earnings difference between parents are similar across the three models studied in this paper, their implications for policy are vastly different. The estimated gap from the standard event-study model mostly driven by negative effects on maternal earnings, while the estimated gap in the event-IV model is driven by both the positive effect estimates for partners and the negative effect on women in equal parts. This shows that the interpretation of the child penalty may not always be as straightforward as commonly believed.

The new insights in the nature of selection into fertility brought forward in this paper show that common intuitions regarding parallel-trend assumptions can be misleading, and that pre-trends are uninformative about the sign of the selection bias in the treatment period. We think of this as a cautionary tale for event-study designs more generally, as it draws attention to the importance of understanding selection from a dynamic rather than a static point of view.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics 113*(2), 231–263.

Adda, J., C. Dustmann, and K. Stevens (2017). The career costs of children. *Journal of Political Economy 125*(2), 293–337.

Aguero, J. M. and M. S. Marks (2008, May). Motherhood and Female Labor Force Participation: Evidence from Infertility Shocks. *American Economic Review 98*(2), 500–504.

Anderson, D. J., M. Binder, and K. Krause (2003). The motherhood wage penalty revisited: Experience, heterogeneity, work effort, and work-schedule flexibility. *Industrial and Labor Relations Review 56*(2), 273–294.

Andresen, M. E. and T. Havnes (2019). Child care, parental labor supply and tax revenue. *Labour Economics 61*, 101762.

Andresen, M. E. and E. Nix (2022). What Causes the Child Penalty? Evidence from Adopting and Same-Sex Couples. *Journal of Labor Economics 40*(4), 971–1004.

Angelov, N., P. Johansson, and E. Lindahl (2016). Parenthood and the gender gap in pay. *Journal of Labor Economics 34*(3), 545–579.

Angrist, J. D. and W. N. Evans (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review 88*(3), 450–477.

Angrist, J. D., B. Ferman, C. Gao, P. Hull, O. L. Tecchio, and R. W. Yeh (2025). Instrumental variables with time-varying exposure: New estimates of revascularization effects on quality of life.

Bhalotra, S. R. and D. Clarke (2019). Twin Births and Maternal Condition. *Review of Economics and Statistics 101*(5), 853–864.

Bhalotra, S. R., D. Clarke, H. Mühlrad, and M. Palme (2019). Multiple Births, Birth Quality and Maternal Labor Supply: Analysis of IVF Reform in Sweden. IZA Discussion Paper No. 12490, IZA Institute of Labor Economics.

Bögl, S., J. Moshfegh, P. Persson, and M. Polyakova (2024). The economics of infertility: Evidence from reproductive medicine. Technical report, National Bureau of Economic Research.

Borusyak, K., X. Jaravel, and J. Spiess (2024, 02). Revisiting event-study designs: Robust and efficient estimation. *The Review of Economic Studies 0*, 1–33.

Bronars, S. G. and J. Grogger (1994). The economic consequences of unwed motherhood: Using twin births as a natural experiment. *American Economic Review 84*(5), 1141–1156.

Callaway, B. and P. H. C. Sant'Anna (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

CDC (2012). Assisted Reproductive Technology, National Summary. Technical report, Center for Disease Control, Atlanta.

Cristia, J. P. (2008). The effect of a first child on female labor supply evidence from women seeking fertility services. *Journal of Human Resources 43*(3), 487–510.

de Chaisemartin, C. and X. D'Haultfœuille (2020, September). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review 110*(9), 2964–96.

Drange, N. and T. Havnes (2019). Child care before age two and the development of language and numeracy: Evidence from a lottery. *Journal of Labor Economics 37*(2), 581–620.

Ferman, B. and O. Tecchio (2025). Dynamic lates with a static instrument.

Gallen, Y., J. S. Joensen, E. R. Johansen, and G. F. Veramendi (2023). The la-

bor market returns to delaying pregnancy. Technical report, Available at SSRN 4554407.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277.

Groes, F., D. Iorio, M. Y. Leung, and R. Santaeulàlia-Llopis (2017). Educational Disparities in the Battle Against Infertility: Evidence from IVF Success. Working paper: 977, Barcelona School of Economics.

Heckman, J. and T. McCurdy (1980). A life cycle model of female labour supply. *Review of Economic Studies 47*(1), 47–74.

Hotz, V. J., S. W. McElroy, and S. G. Sanders (2005). Teenage childbearing and its life cycle consequences exploiting a natural experiment. *Journal of Human Resources 40*(3), 683–715.

Hotz, V. J. and R. A. Miller (1988). An empirical analysis of life cycle fertility and female labor supply. *Econometrica 56*(1), 91–118.

Kleven, H. (2022). The geography of child penalties and gender norms: Evidence from the United States. Working paper 30176, National Bureau of Economic Research.

Kleven, H., C. Landais, J. Posch, A. Steinhauer, and J. Zweimüller (2019, May). Child penalties across countries: Evidence and explanations. *AEA Papers and Proceedings 109*, 122–26.

Kleven, H., C. Landais, and J. E. Søgaard (2019). Children and gender inequality: Evidence from Denmark. *American Economic Journal: Applied Economics 11*(4), 181–209.

Korenman, S. and D. Neumark (1992). Marriage, Motherhood, and Wages. *The Journal of Human Resources 27*(2), 233–255.

Lundborg, P., E. Plug, and A. W. Rasmussen (2017). Can women have children and a career? IV evidence from IVF treatments. *American Economic Review 107*(6), 1611–1637.

Lundborg, P., E. Plug, and A. W. Rasmussen (2024). Is there really a child penalty in the long run? New evidence from IVF treatments. IZA Discussion Paper No. 16959, IZA Institute of Labor Economics.

Magnus, P., L. M. Irgens, K. Haug, W. Nystad, R. Skjærven, and C. Stoltenberg (2006). Cohort profile: The Norwegian mother and child cohort study (MoBa). *International Journal of Epidemiology 35*(5), 1146–1150.

Martinenghi, F. and M. Naghsh-Nejad (2025). Career, family, and ivf: The impact of involuntary childlessness and fertility treatment. IZA Discussion Paper No.17965, IZA Institute of Labor Economics.

Miller, A. R. (2011). The effects of motherhood timing on career path. *Journal of Population Economics 24*(3), 1071–1100.

NOU 2017:6 (2017). Offentlig støtte til barnefamiliene. Technical report, Ministry of Children and Families.

Rambachan, A. and J. Roth (2023, 02). A more credible approach to parallel trends. *The Review of Economic Studies 90*(5), 2555–2591.

Rosenzweig, M. R. and K. I. Wolpin (1980). Life-cycle labor supply and fertility: Causal inferences from household models. *Journal of Political Economy 88*(2), 328–348.

Roth, J., P. H. Sant'Anna, A. Bilinski, and J. Poe (2023). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics 235*(2), 2218–2244.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199.

Waldfogel, J. (1997). The effect of children on women's wages. *American Sociological Review 62*(2), 209–217.

A Appendix For Online Publication

A.1 Mapping LPR-IV to child-age

Let $\text{success}_i \in \{0,1\}$ indicate whether the first IVF leads to a birth roughly nine months later. For $s \in \{0,1\}$, let $T_i(s)$ be the (calendar) year of first birth if $\text{success}_i = s$ (with $T_i(s) = \infty$ if no birth), and define potential time-to-birth since IVF

$$P_i(s) \equiv T_i(s) - IVF_i \in \{0,1,2,\ldots,\infty\}.$$

Timing is defined such that the first IVF attempt either results in a birth in the first post-IVF period (so $P_i(1) = 0$ for all i), or it does not (so $P_i(0) > 0$); any births after a failure arise in later periods via subsequent IVF or natural conception.

Exclusion assumes that conditional on child age $a \in \{\infty, 0, 1, \ldots\}$, potential outcomes do not depend on IVF success $s \in \{0,1\}$

$$y_{it}^{a,1} = y_{it}^{a,0} = y_{it}^{a}, \quad \forall a$$

Exogeneity assumes that, conditional on women's age, education and timing of the IVF, the result of the IVF attempt is random, and therefore independent of potential outcomes y_{it}^{a} and potential counterfactual fertility $P_i(0)$:

$$\{y_{it}^{a}\}_{a\in\{\infty,0,1,\ldots\}}, P_i(0) \perp\!\!\!\perp \text{success}_i \mid age_i, edu_i, IVF_i$$

We document in the paper that the data are consistent with this assumption.

By definition of the instrument a successful IVF implies a birth in $p = 0$: $P_i(1) \equiv T_i(1) - IVF_i = 0$, and timing implies $P_i(0) > 0$.

Monotonicity,

$$0 = P_i(1) \leq P_i(0)$$

is therefore mechanically satisfied. Instrument *relevance* can be checked in the data (see Figure 1).

By definition,

$$a_{ip}(s) = p - P_i(s) \quad \text{with} \quad a_{ip}(s) = \infty \text{ if } P_i(s) > p,\ s \in \{0,1\}.$$

and the age-horizon specific first-stage equals:

$$\begin{aligned}\pi_{ap} &\equiv E\big[\mathbf{1}\{a_{ip} = a\} \mid \text{success}_i = 1\big] - E\big[\mathbf{1}\{a_{ip} = a\} \mid \text{success}_i = 0\big] \\ &= \mathbf{1}\{a = p\} - \Pr\big(P_i(0) = p - a\big).\end{aligned}$$

since

$$E\big[\mathbf{1}\{a_{ip}=a\} \mid \text{success}_i=1\big] = \Pr\big(a_{ip}(1)=a\big) = \Pr\big(P_i(1)=p-a\big) = \mathbf{1}\{a=p\},$$

because $P_i(1)=0$ for all i, and

$$E\big[\mathbf{1}\{a_{ip}=a\} \mid \text{success}_i=0\big] = \Pr\big(a_{ip}(0)=a\big) = \Pr\big(P_i(0)=p-a\big).$$

Fertility p years since IVF is defined as

$$\text{Fertility}_{ip} = \mathbf{1}\{P_i \le p\} = \sum_{a=0}^{p} \mathbf{1}\{a_{ip}=a\},$$

since $\mathbf{1}\{a_{ip}=a\}=0$ for all $a>p$. Summing over a gives the fertility first stage

$$\begin{aligned}
\pi_p &= E\big[\text{Fertility}_{ip} \mid \text{success}_i=1\big] - E\big[\text{Fertility}_{ip} \mid \text{success}_i=0\big] \\
&= E\Big[\sum_{a=0}^{p} \mathbf{1}\{a_{ip}=a\} \mid \text{success}_i=1\Big] - E\Big[\sum_{a=0}^{p} \mathbf{1}\{a_{ip}=a\} \mid \text{success}_i=0\Big] \\
&= \sum_{a=0}^{p} \big(E\big[\mathbf{1}\{a_{ip}=a\} \mid \text{success}_i=1\big] - E\big[\mathbf{1}\{a_{ip}=a\} \mid \text{success}_i=0\big]\big) \\
&= \sum_{a=0}^{p} \pi_{ap}.
\end{aligned}$$

To recover the reduced form, fix $p \ge 0$. Let $p = t - IVF_i$ then

$$y_{ip}(s) = y_{ip}^{\infty}\, \mathbf{1}\{P_i(s) > p\} + \sum_{a=0}^{p} y_{ip}^{a}\, \mathbf{1}\{P_i(s)=p-a\}, \qquad s \in \{0,1\}.$$

Since $P_i(1)=0$ and $P_i(0)>0$, we have $y_{ip}(1) = y_{ip}^{p}$. Under failure,

$$y_{ip}(0) = y_{ip}^{\infty}\, \mathbf{1}\{P_i(0) > p\} + \sum_{a=0}^{p-1} y_{ip}^{a}\, \mathbf{1}\{P_i(0)=p-a\}.$$

Therefore the reduced form

$$\rho_p \equiv E[y_{ip} \mid \text{success}_i=1] - E[y_{ip} \mid \text{success}_i=0]$$

equals

$$\begin{aligned}
\rho_p &= E[y_{ip}^p] - E\Big[y_{ip}^{\infty}\,\mathbf{1}\{P_i(0) > p\} + \sum_{a=0}^{p-1} y_{ip}^a\,\mathbf{1}\{P_i(0) = p-a\}\Big] \\
&= E\big[y_{ip}^p - y_{ip}^{\infty}\big] - \sum_{a=0}^{p-1} E\big[(y_{ip}^a - y_{ip}^{\infty})\,\mathbf{1}\{P_i(0) = p-a\}\big], \\
&= \underbrace{E[y_{ip}^p - y_{ip}^{\infty}]}_{\delta_{pp}} + \sum_{a=0}^{p-1} \underbrace{E[y_{ip}^a - y_{ip}^{\infty} \mid P_i(0) = p-a]}_{\delta_{pa}} \underbrace{\big(-\Pr(P_i(0) = p-a)\big)}_{\pi_{ap}} \\
&= \sum_{a=0}^{p} \delta_{ap}\pi_{ap}
\end{aligned}$$

This shows that

$$\gamma_p = \frac{\rho_p}{\pi_p} = \frac{\sum_{a=0}^{p} \pi_{ap}\,\delta_{ap}}{\sum_{a=0}^{p} \pi_{ap}} = \sum_{a=0}^{p} \omega_{ap}\,\delta_{ap}, \qquad \omega_{ap} \equiv \frac{\pi_{ap}}{\sum_{a'=0}^{p} \pi_{a'p}}. \tag{13}$$

Thus, γ_p is a first-stage–weighted average of child-age effects for $a \in \{0,\ldots,p\}$, where

$$\delta_{ap} = \begin{cases} E[y_{ip}^a - y_{ip}^{\infty}] & a = p \\ E[y_{ip}^a - y_{ip}^{\infty} \mid P_i(0) = p-a] & a < p \end{cases}$$

$$\pi_{ap} = \begin{cases} 1 & a = p \\ -\Pr\big(P_i(0) = p-a\big) & a < p \end{cases}$$

Note that $\sum_a \omega_{ap} = 1$ but weights for $a < p$ can be negative because $\pi_{ap} = -\Pr(P_i(0) = p-a)$.

A.2 Standard errors on rescaled estimates

Denote the rescaled estimate by x:

$$x = \frac{y^1 - y^{\infty}}{y^{\infty}} \equiv \frac{\delta}{y^{\infty}}$$

The Delta method gives

$$V(x) = \begin{pmatrix} \partial x/\partial \delta \\ \partial x/\partial y^{\infty} \end{pmatrix}' V \begin{pmatrix} \delta \\ y^{\infty} \end{pmatrix} \begin{pmatrix} \partial x/\partial \delta \\ \partial x/\partial y^{\infty} \end{pmatrix}$$

where

$$V\begin{pmatrix}\delta\\ y^{\infty}\end{pmatrix}=\begin{pmatrix}V(\delta) & cov(\delta,y^{\infty})\\ & V(y^{\infty})\end{pmatrix}=\begin{pmatrix}V(\delta) & (V(y^{1})-V(y^{\infty})-V(\delta))/2\\ & V(y^{\infty})\end{pmatrix}$$

since

$$\begin{aligned}V(\delta)&=V(y^{1})+V(y^{\infty})-2cov(y^{1},y^{\infty})\\ \Rightarrow cov(y^{1},y^{\infty})&=(V(y^{1})+V(y^{\infty})-V(\delta))/2\\ &\text{from this we get}\\ cov(\delta,y^{\infty})&=cov(y^{1},y^{\infty})-V(y^{\infty})\\ &=(V(y^{1})-V(y^{\infty})-V(\delta))/2\end{aligned}$$

we also have that

$$\begin{pmatrix}\partial x/\partial\delta\\ \partial x/\partial y^{\infty}\end{pmatrix}=\begin{pmatrix}1/y^{\infty}\\ -x/y^{\infty}\end{pmatrix}$$

which implies that the variance on the rescaled estimate is as follows

$$\begin{aligned}V(x)&=(V(\delta)-2\cdot x\cdot cov(\delta,y^{\infty})+x^{2}V(y^{\infty}))/(y^{\infty})^{2}\\ &=(V(\delta)-x\cdot(V(y^{1})-V(y^{\infty})-V(\delta))+x^{2}V(y^{\infty}))/(y^{\infty})^{2}\end{aligned}$$

where $V(\delta), V(y^{1})$ and $V(y^{\infty})$, all come from separate 2SLS regressions as outlined in section 4.4.

A.3 Additional descriptive statistics

Table A1. Descriptive statistics for IVF women by success at first trial

	(1) Failure	(2) Success	(3) Difference	
Woman characteristics				
Number of IVF attempts	3.31	1.81	-1.49	(0.03)
Success, endpoint	0.46	1.00	0.54	(0.01)
Total number of children	1.31	1.84	0.54	(0.02)
0 children	0.24	0.00	0.24	(0.00)
1 children	0.29	0.30	0.00	(0.01)
2 children	0.38	0.58	0.20	(0.01)
3 children	0.08	0.12	0.04	(0.01)
4 children	0.01	0.01	0.00	(0.00)
Age	32.1	31.3	-0.79	(0.09)
Education				
- Compulsory	0.15	0.12	-0.03	(0.01)
- High School	0.24	0.23	-0.01	(0.01)
- Bachelor	0.41	0.44	0.03	(0.01)
- Master	0.20	0.21	0.01	(0.01)
Earnings (1000 NOK)	362.8	362.6	-0.13	(4.14)
Hours (FTE)	0.88	0.88	0.00	(0.01)
Employed	0.80	0.81	0.01	(0.01)
Hourly wage (NOK)	221.5	220.4	-1.07	(4.44)
Sickness absence days	15.1	14.7	-0.42	(0.73)
Visits to general practitioner	2.53	2.47	-0.06	(0.05)
Psychological symptoms	0.14	0.14	-0.00	(0.01)
Hospital days	2.21	1.95	-0.27	(0.15)
Partner characteristics				
Age	35.3	34.5	-0.83	(0.13)
Female	0.01	0.02	-0.01	(0.01)
Education				
- Compulsory	0.17	0.16	-0.02	(0.01)
- High School	0.39	0.37	0.03	(0.01)
- Bachelor	0.27	0.29	0.01	(0.01)
- Master	0.17	0.18	-0.63	(6.36)
Earnings (1000 NOK)	455.1	454.4	0.00	(0.01)
Hours (FTE)	0.84	0.84	0.00	(0.00)
Employed	0.83	0.84	0.01	(0.01)
Hourly wage (NOK)	280.7	282.4	1.71	(5.12)
N Women	6 881	3 152		

Notes: Table shows mean characteristics of all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time. The statistics are presented by success and failure at first trial, as well as the difference (with standard error) between the two. Labor market outcomes and health indicators are measured as averages over the four years prior to the first IVF trial. Age and education are measured the year before the IVF treatment.

Table A2. Total number of children.

	(1)	(2)	(3)
Number of children	Non-IVF sample	Reweighted non-IVF sample	IVF sample
1	0.23	0.34	0.36
2	0.60	0.55	0.53
3	0.16	0.10	0.11
≥ 4	0.02	0.01	0.01
Number of women	109,791	109,791	8,346

Note: This table shows the number of children by the end of the sample period, conditional on having at least one child. Column 1 shows fertility for the non-IVF sample; column 2 for the non-IVF sample reweighted to match the distribution of the IVF sample; and column 3 for the subsample of the IVF sample which includes women with at least one child.

Table A3. Descriptive statistics for IVF women vs reweighted non-IVF women

	(1) IVF	(2) Non-IVF (reweighted)	(3) Difference	
Woman characteristics				
Number of IVF attempts	2.84			
Success, first attempt	0.31			
Success, endpoint	0.63			
Total number of children	1.47	1.79	-0.31	(0.01)
0 children	0.17			
1 child	0.30	0.34	-0.04	(0.00)
2 children	0.44	0.55	-0.11	(0.00)
3 children	0.09	0.10	-0.02	(0.00)
4 children	0.01	0.01	-0.00	(0.00)
Age	31.8	31.8	-0.00	(0.04)
Education				
- Compulsory	0.14	0.14	0.00	(0.00)
- High School	0.24	0.24	-0.00	(0.00)
- Bachelor	0.42	0.42	-0.00	(0.00)
- Master	0.20	0.20	0.00	(0.00)
Yearly earnings (1000 NOK)	362.7	348.3	14.4	(1.80)
Hours (FTE)	0.88	0.85	0.03	(0.00)
Employed	0.80	0.73	0.07	(0.00)
Hourly earnings (NOK)	221.1	219.8	1.30	(1.84)
Sickness absence days	15.0	12.8	2.17	(0.31)
Visits to general practitioner	2.51	2.12	0.39	(0.02)
Psychological symptoms	0.14	0.13	0.01	(0.00)
Hospital days	2.13	0.99	1.13	(0.08)
Partner characteristics				
Age	35.1	34.2	0.85	(0.06)
Female	0.01	0.01	0.00	(0.00)
Education				
- Compulsory	0.17	0.17	-0.00	(0.00)
- High School	0.39	0.35	0.04	(0.00)
- Bachelor	0.27	0.29	-0.01	(0.00)
- Master	0.17	0.19	-0.02	(0.00)
Earnings (1000 NOK)	457.1	425.5	31.53	(2.82)
Hours (FTE)	0.85	0.80	0.05	(0.00)
Employed	0.84	0.77	0.06	(0.00)
Hourly earnings (NOK)	281.2	276.5	4.73	(2.09)
Observations	10 033	108 786		

Notes: Table shows mean characteristics for the IVF sample and the reweighted non-IVF sample, as well as the difference (with standard error) between the two. By construction, the non-IVF sample includes only women with at least one child. Labor market outcomes and health indicators are measured as averages over the four years prior to the first IVF trial, or, for non-IVF women, prior to the approximate conception date. Age and education are measured the year before the IVF treatment.

A.4 Main estimates in table form

Table A4. Point estimates from the event study model.

Age of child	Woman (1)	Partner (2)	Gap (1) - (2)	Age of child	Woman (3)	Partner (4)	Gap (3) - (4)
-12	-0.266	-0.342	0.076	0	-0.169	-0.003	-0.165
	(0.034)	(0.031)	(0.038)		(0.004)	(0.005)	(0.006)
-11	-0.237	-0.273	0.036	1	-0.266	-0.043	-0.222
	(0.030)	(0.030)	(0.037)		(0.005)	(0.006)	(0.007)
-10	-0.211	-0.257	0.046	2	-0.174	-0.015	-0.159
	(0.027)	(0.026)	(0.033)		(0.007)	(0.010)	(0.010)
-9	-0.195	-0.215	0.020	3	-0.171	-0.007	-0.165
	(0.022)	(0.024)	(0.030)		(0.008)	(0.011)	(0.013)
-8	-0.154	-0.164	0.011	4	-0.164	0.001	-0.166
	(0.020)	(0.022)	(0.027)		(0.010)	(0.014)	(0.015)
-7	-0.117	-0.140	0.023	5	-0.146	0.013	-0.159
	(0.018)	(0.018)	(0.024)		(0.012)	(0.017)	(0.018)
-6	-0.082	-0.112	0.030	6	-0.130	0.024	-0.153
	(0.015)	(0.016)	(0.021)		(0.014)	(0.019)	(0.021)
-5	-0.053	-0.087	0.034	7	-0.128	0.032	-0.160
	(0.013)	(0.013)	(0.016)		(0.015)	(0.022)	(0.024)
-4	-0.025	-0.053	0.029	8	-0.129	0.042	-0.171
	(0.010)	(0.010)	(0.013)		(0.017)	(0.026)	(0.028)
-3	0.001	-0.031	0.032	9	-0.122	0.044	-0.166
	(0.007)	(0.007)	(0.009)		(0.019)	(0.031)	(0.033)
-2	0.008	-0.017	0.025	10	-0.119	0.029	-0.148
	(0.004)	(0.004)	(0.006)		(0.022)	(0.034)	(0.036)
				11	-0.127	0.032	-0.158
					(0.024)	(0.040)	(0.044)
				12	-0.134	-0.016	-0.118
					(0.033)	(0.041)	(0.049)
χ^2-test on pre	198.80	165.31	42.81				
p-val.	0.00	0.00	0.00				

Note: Table shows point estimates and standard errors for the event study model and are equivalent to estimates presented in Figure 3. The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were registered with a partner at the time, and eventually had at least one child (observations = 145,571; unique women = 8,349).

Table A5. Point estimates from the LPR-IV model.

Time since IVF	Woman (1)	Partner (2)	Gap (1) - (2)	Time since IVF	Woman (3)	Partner (4)	Gap (3) - (4)
-12				0	-0.169	0.021	-0.190
					(0.014)	(0.019)	(0.023)
-11				1	-0.292	-0.033	-0.259
					(0.025)	(0.032)	(0.040)
-10				2	-0.008	0.070	-0.078
					(0.038)	(0.047)	(0.055)
-9				3	-0.061	0.113	-0.174
					(0.045)	(0.057)	(0.065)
-8				4	-0.053	0.141	-0.194
					(0.049)	(0.070)	(0.078)
-7				5	-0.015	0.165	-0.180
					(0.056)	(0.078)	(0.087)
-6				6	0.002	0.156	-0.154
					(0.067)	(0.084)	(0.100)
-5				7	-0.047	0.216	-0.263
					(0.074)	(0.098)	(0.116)
-4				8	-0.046	0.171	-0.218
					(0.082)	(0.145)	(0.159)
-3				9	-0.004	0.204	-0.208
					(0.096)	(0.147)	(0.166)
-2				10	0.063	-0.028	0.091
					(0.104)	(0.160)	(0.171)
				11	0.066	0.108	-0.042
					(0.118)	(0.174)	(0.197)
				12	0.197	-0.088	0.285
					(0.166)	(0.223)	(0.263)

Note: Table shows point estimates and standard errors for the LPR-IV model and are equivalent to estimates presented in Figure 4. The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033).

Table A6. Point estimates from the Event-IV model.

	Woman	Partner	Gap		Woman	Partner	Gap
Age of child	(1)	(2)	(1) - (2)	Age of child	(3)	(4)	(3) - (4)
-12	0.150	-0.083	0.233	0	0.052	0.029	0.024
	(0.445)	(0.338)	(0.542)		(0.042)	(0.053)	(0.062)
-11	0.123	-0.029	0.152	1	-0.165	0.021	-0.186
	(0.344)	(0.437)	(0.512)		(0.014)	(0.019)	(0.022)
-10	0.110	-0.110	0.220	2	-0.238	-0.014	-0.224
	(0.242)	(0.244)	(0.312)		(0.017)	(0.025)	(0.030)
-9	0.079	-0.101	0.180	3	-0.121	0.025	-0.146
	(0.163)	(0.183)	(0.218)		(0.022)	(0.031)	(0.035)
-8	0.054	-0.056	0.109	4	-0.114	0.047	-0.161
	(0.127)	(0.146)	(0.171)		(0.025)	(0.036)	(0.039)
-7	0.012	-0.022	0.034	5	-0.101	0.065	-0.166
	(0.096)	(0.122)	(0.138)		(0.027)	(0.042)	(0.045)
-6	0.009	-0.020	0.029	6	-0.081	0.080	-0.162
	(0.075)	(0.101)	(0.113)		(0.029)	(0.046)	(0.049)
-5	0.029	-0.005	0.033	7	-0.067	0.087	-0.153
	(0.062)	(0.085)	(0.094)		(0.033)	(0.050)	(0.055)
-4	0.053	0.010	0.043	8	-0.069	0.110	-0.180
	(0.055)	(0.072)	(0.079)		(0.036)	(0.055)	(0.061)
-3	0.052	0.017	0.035	9	-0.067	0.110	-0.176
	(0.049)	(0.063)	(0.070)		(0.039)	(0.069)	(0.074)
-2	0.047	0.027	0.020	10	-0.053	0.125	-0.178
	(0.045)	(0.058)	(0.066)		(0.045)	(0.077)	(0.084)
				11	-0.030	0.069	-0.100
					(0.050)	(0.077)	(0.083)
				12	-0.018	0.091	-0.109
					(0.055)	(0.085)	(0.095)
χ^2-test on pre	8.54	10.64	9.19				
p-val.	0.66	0.47	0.60				

Note: Table shows point estimates and standard errors for the IV model and are equivalent to estimates presented in Figure 5. The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033)

A.5 LPR-IV First Stage

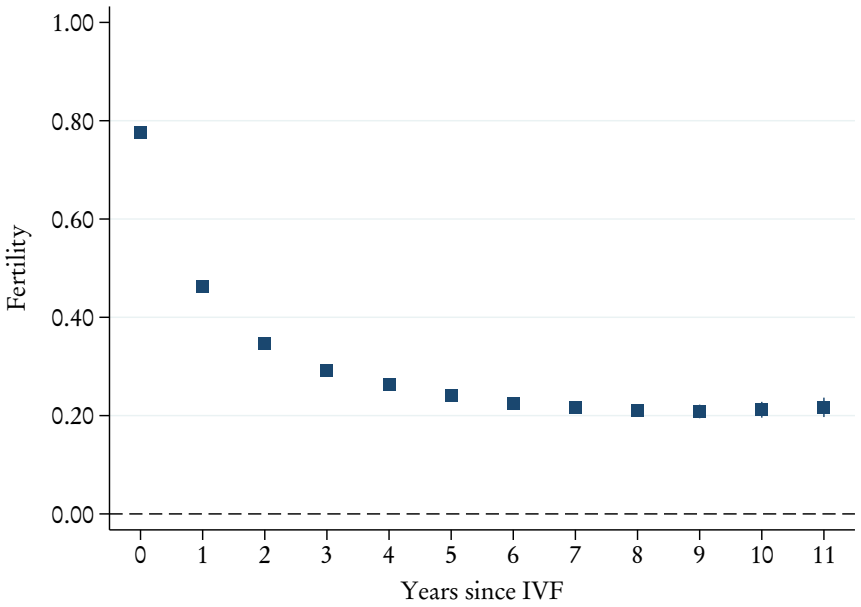
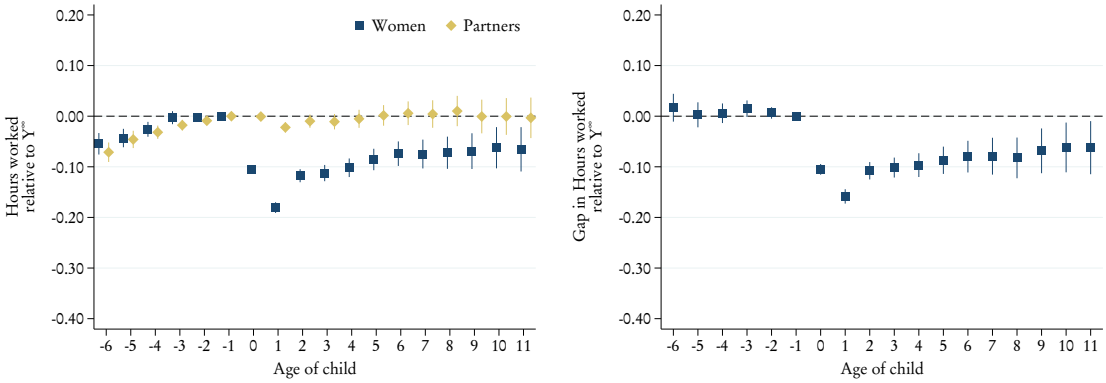


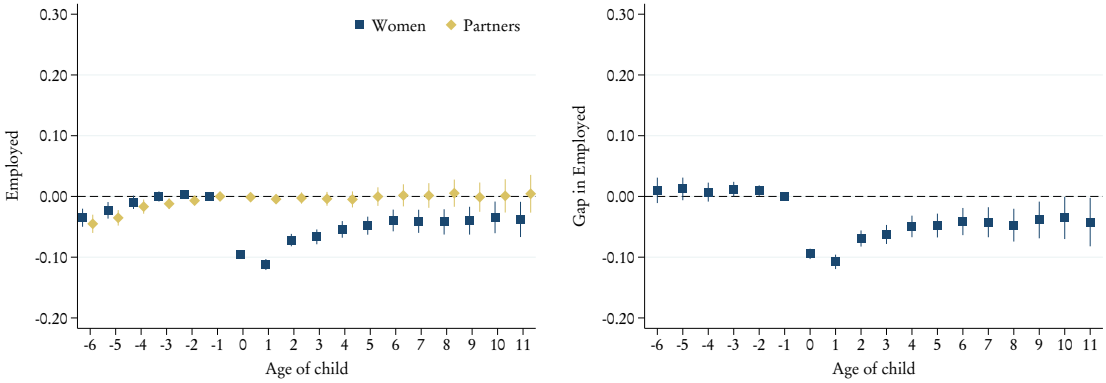
Figure A1. First stage. LPR-IV.

Note: First stage estimates using the IV model of Lundborg et al. (2017) as described in equation (4) on our data. The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033)

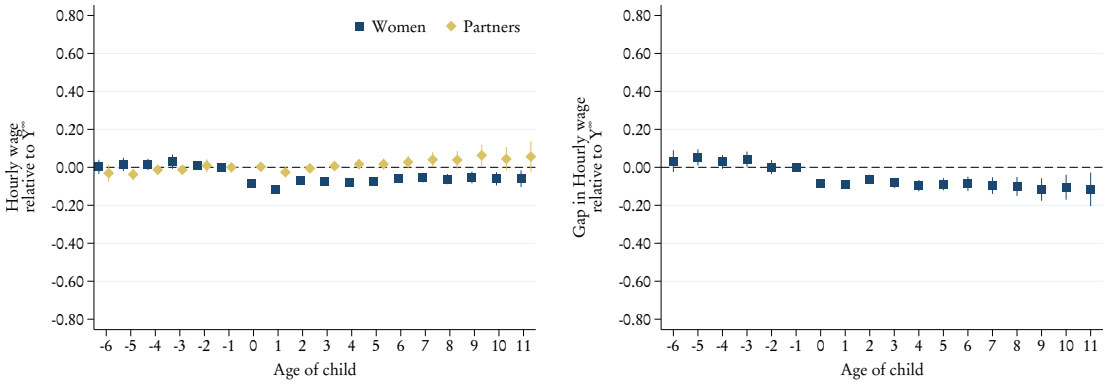
A.6 Other labor market outcomes



(a) Hours worked **(b)** Gap in hours worked



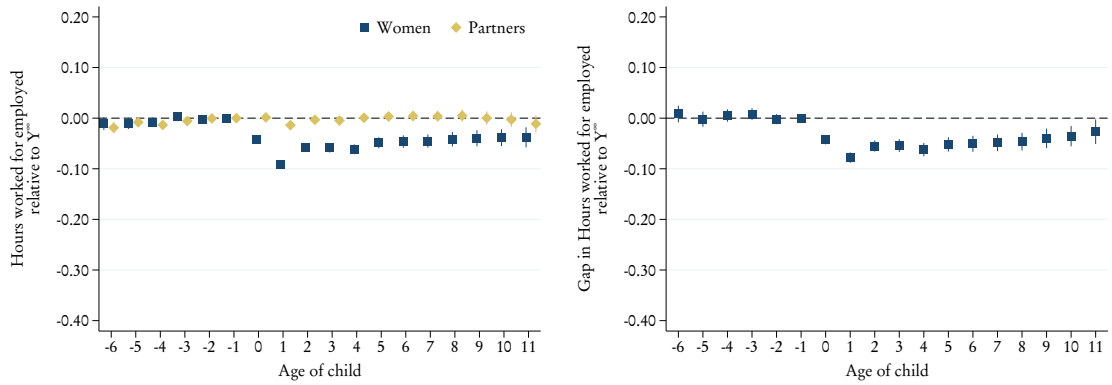
(c) Employment **(d)** Gap in employment



(e) Hourly earnings **(f)** Gap in hourly earnings

Figure A2. Other labor market outcomes. Event.

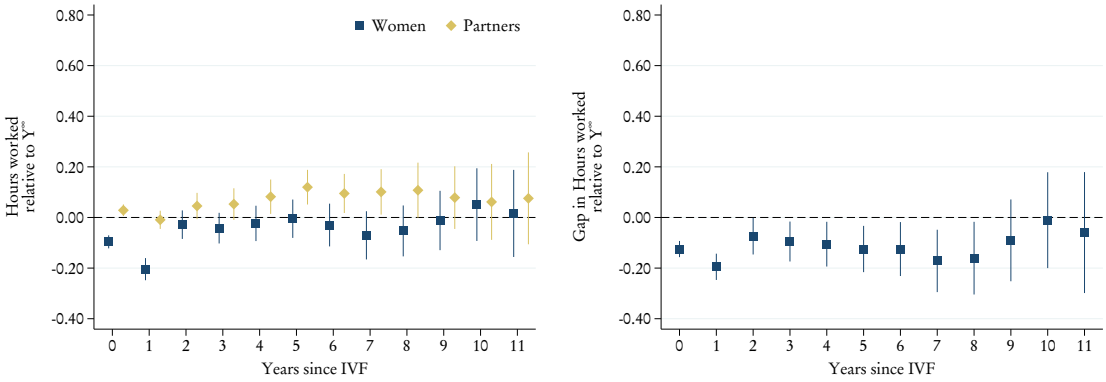
Note: Event-study estimates from specification (2). Outcomes are employment (panel a and b), hours worked (panel c and d), and hourly wages (panel e and f). Panel a, c, and e show effects separately for women and partners, figures b, d, and f show difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 4.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were registered with a partner at the time, and eventually had at least one child (observations = 145,571; unique women = 8,349).



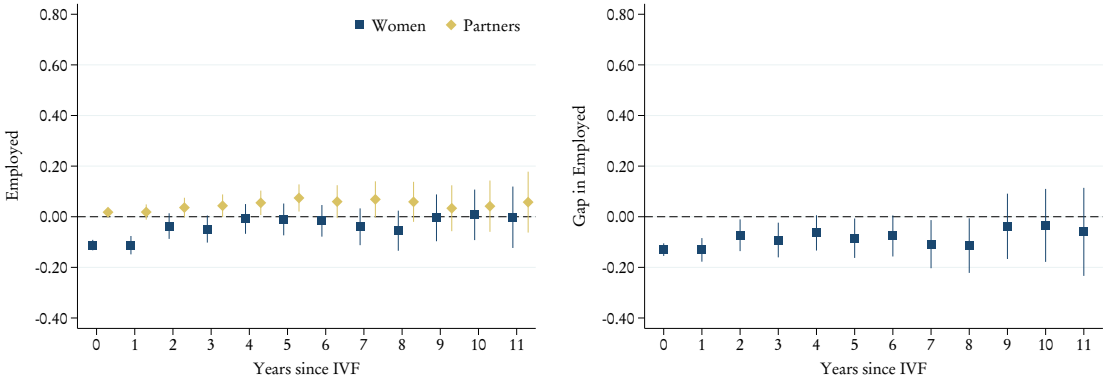
(a) Hours worked for employed **(b)** Gap in hours worked for employed

Figure A3. Hours conditional on employment. Event.

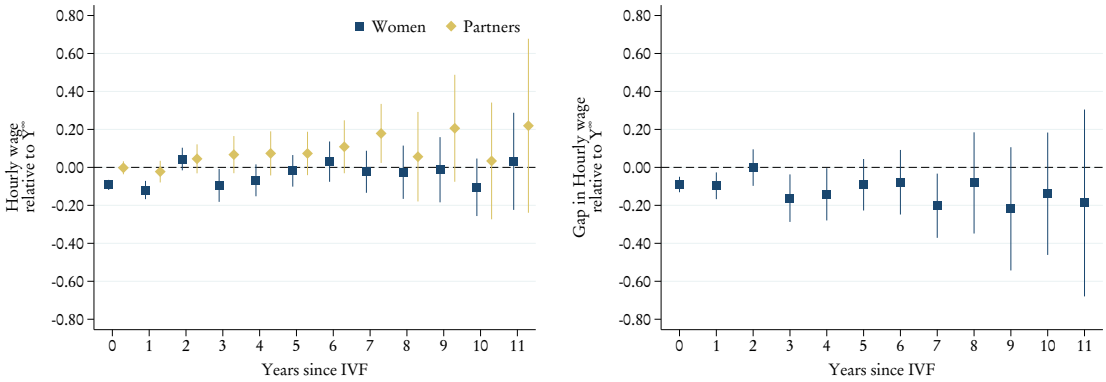
Note: Event-study estimates from specification (2). Outcome is hours worked conditional on employment. Panel a shows effects separately for women and partners, panel b shows difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 4.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were registered with a partner at the time, and eventually had at least one child (observations = 145,571; unique women = 8,349).



(a) Hours worked **(b)** Gap in hours worked



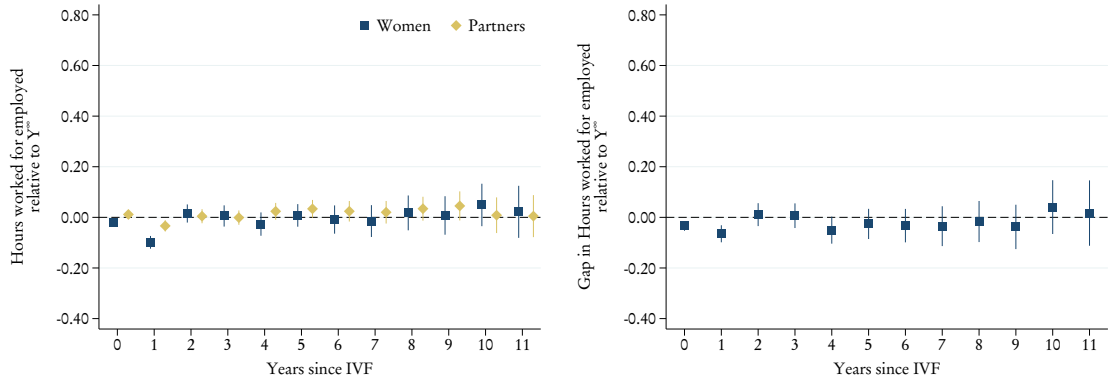
(c) Employment **(d)** Gap in employment



(e) Hourly earnings **(f)** Gap in hourly earnings

Figure A4. Other labor market outcomes. LPR-IV.

Note: Estimated effects of fertility using the IV model of Lundborg et al. (2017) as described in equation (4) on our data. Outcomes are employment (panel a and b), hours worked (panel c and d), and hourly wages (panel e and f). Panels a, c, and e show effects separately for women and partners, panels b, d, and f show difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 4.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033)

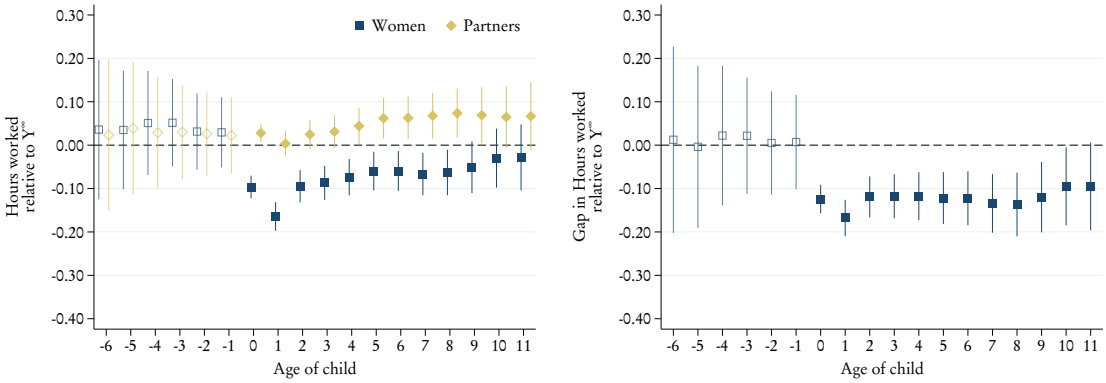


(a) Hours worked for employed

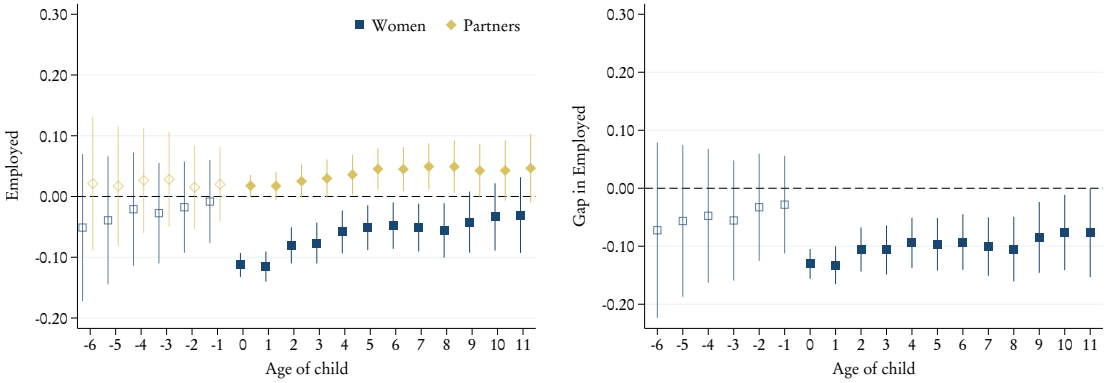
(b) Gap in hours worked for employed

Figure A5. Hours conditional on employment. LPR-IV.

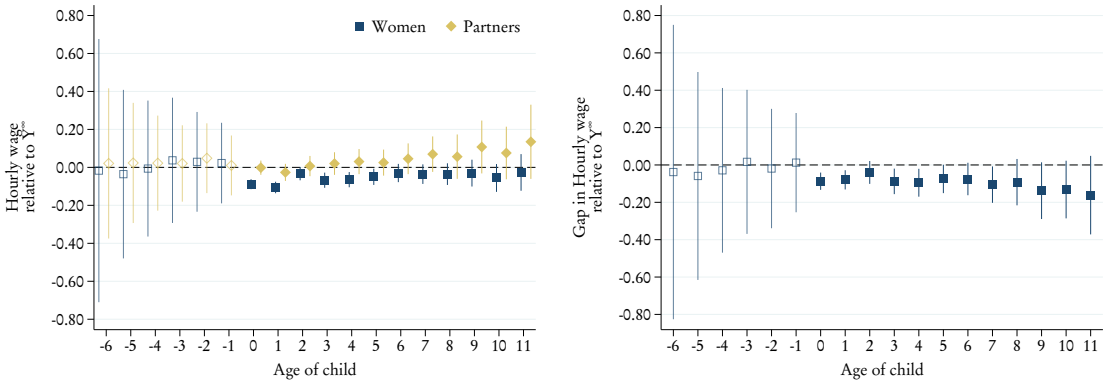
Note: Estimated effects of fertility using the IV model of Lundborg et al. (2017) as described in equation (4) on our data. Outcome is hours worked conditional on employment. Panel a shows effects separately for women and partners, panel b shows difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 4.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033).



(a) Hours worked **(b)** Gap in hours worked



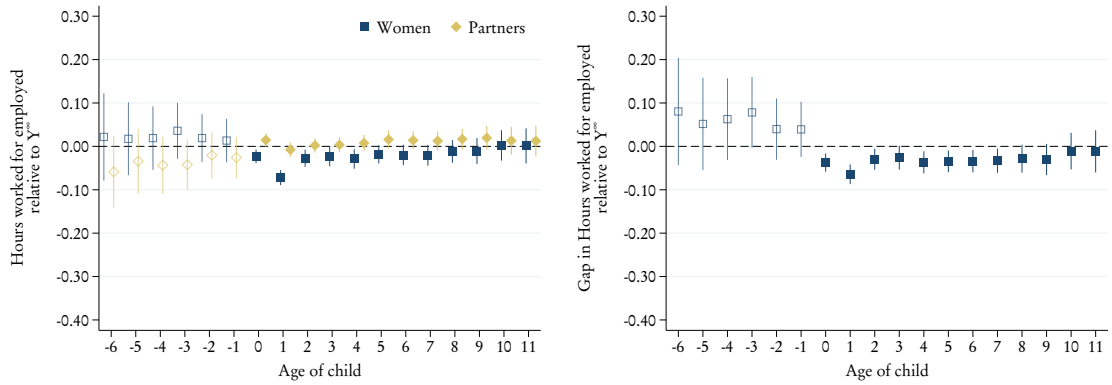
(c) Employment **(d)** Gap in employment



(e) Hourly earnings **(f)** Gap in hourly earnings

Figure A6. Other labor market outcomes. Event-IV.

Note: Estimated effects of age of child using the event-IV model described in equation (10, event-IV). Outcomes are employment (panel a and b), hours worked (panel c and d), and hourly wages (panel e and f). Panels a, c, and e show effects separately for women and partners, panels b, d, and f show difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 4.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033)



(a) Hours worked for employed **(b)** Gap in hours worked for employed

Figure A7. Hours conditional on employment. Event-IV.

Note: Estimated effects of age of child using the event-IV model described in equation (10, event-IV). Outcome is hours worked conditional on employment. Panel a shows effects separately for women and partners, panel b shows difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 4.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033)

A.7 Non-labor market outcomes

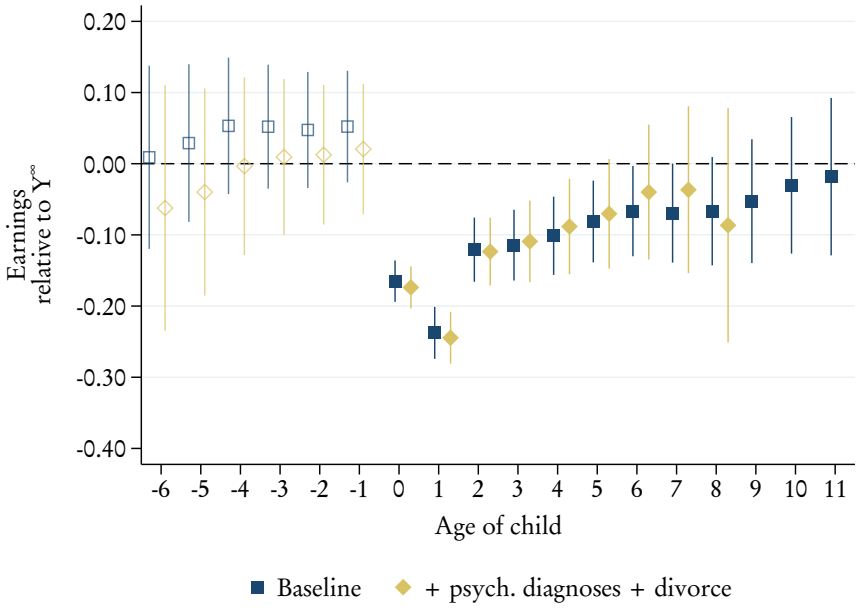
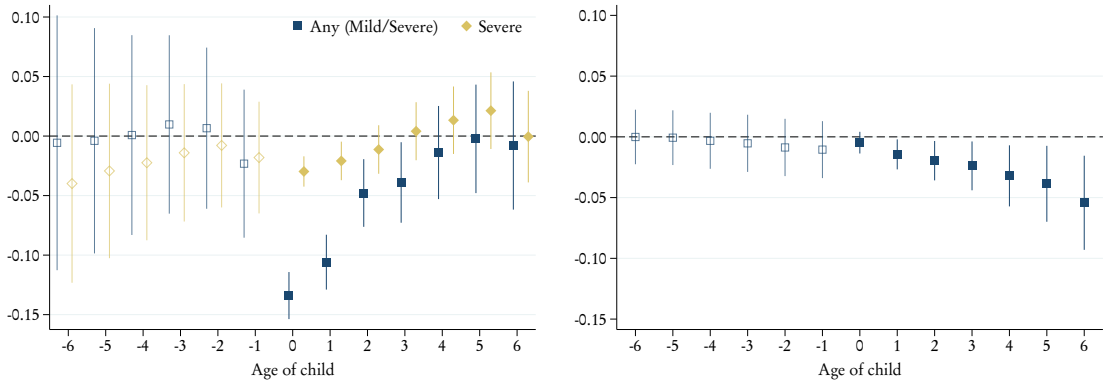


Figure A9. Robustness. Event-IV.

Note: Robustness checks of our event-IV model as specified in equation (10, event-IV). Figure shows our baseline specification, alongside estimates that include controls for divorce and visits to a general practitioner for psychological symptoms. All estimates are scaled relative to counterfactual earnings (Y^{∞}) as described in section 4.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033).



(a) Psychological symptoms **(b)** Divorce

Figure A8. Non-labor market outcomes – Event-IV.

Note: Results from the event-IV model in equation (10, event-IV), using (a) GP visits for psychological symptoms and (b) divorce as outcomes. Panel (a) also reports estimates for severe psychological symptoms. The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033).

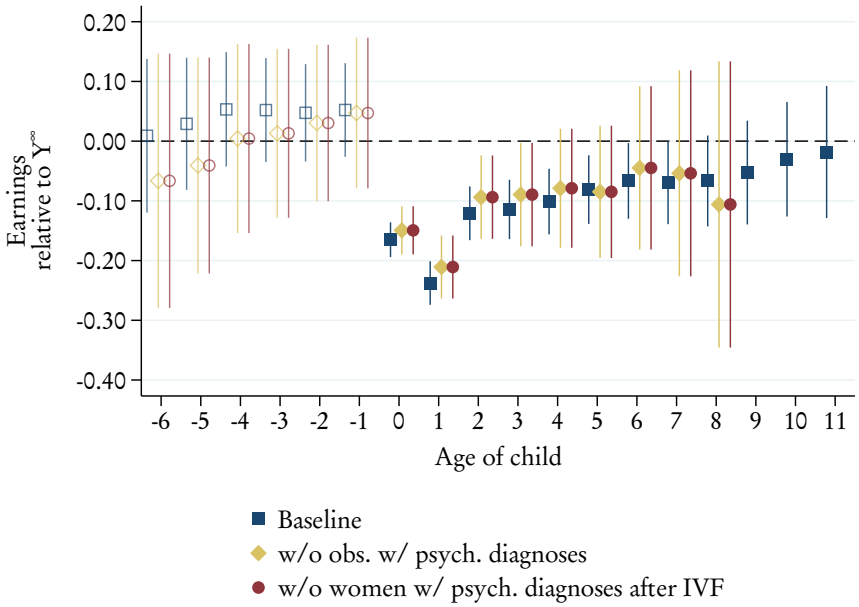


Figure A10. Robustness. Event-IV.

Note: Robustness checks of our event-IV model as specified in equation (10, event-IV). Figure A10 show our baseline specification estimated in the sample of IVF women (all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033)), alongside estimates based on (i) a subsample where we exclude all observations with a psychological diagnosis and (ii) a subsample where we exclude all *women* who received a psychological diagnosis after IVF treatment. All estimates are scaled relative to counterfactual earnings (Y^{∞}) as described in section 4.5.

Table A7. Comparison of long-run (age 11) fertility effects and child penalty estimates across models

	LPR-IV (1)	Event (2)	Event-IV (3)	Event vs. Event-IV (2) - (3)
Gap	-0.04	-0.16	-0.11	-0.05
	(0.22)	(0.04)	(0.10)	(0.10)
Women	0.07	-0.13	-0.02	-0.11
	(0.13)	(0.02)	(0.06)	(0.05)
Partners	0.11	0.03	0.09	-0.06
	(0.18)	(0.04)	(0.08)	(0.09)

Note: Table shows estimates of earnings for women, partners, and the gap (women - partners), evaluated at $a = 11$ ($p = 11$ for LPR-IV). Column (1) shows estimates from the LPR-IV model, column (2) shows the estimates from the event-model, column (3) shows estimates from the event-IV model, and column (2) - (3) shows the difference between the event model and the event-IV model. Standard errors for gaps between parents and differences across models are bootstrapped using 199 repetitions. The sample for the IV estimates includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033). The sample for the event study estimates includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were registered with a partner at the time, and eventually had at least one child (observations = 145,571; unique women = 8,349).

Table A8. Robustness: Comparison of long-run (age 6) fertility effects and child penalty estimates across models, without restriction that woman has partner

	LPR-IV (1)	Event (2)	Event-IV (3)	Event vs. Event-IV (2) - (3)
Gap	-0.07	-0.15	-0.11	-0.04
	(0.10)	(0.02)	(0.06)	(0.06)
Women	0.04	-0.14	-0.05	-0.09
	(0.06)	(0.01)	(0.03)	(0.03)
Partners	0.11	0.01	0.06	-0.05
	(0.10)	(0.02)	(0.06)	(0.06)

Note: Table shows estimates of earnings for women, partners, and the gap (women - partners), evaluated at $a = 6$ ($p = 6$ for LPR-IV). Column (1) shows estimates from the LPR-IV model, column (2) shows the estimates from the event-model, column (3) shows estimates from the event-IV model, and column (2) - (3) shows the difference between the event model and the event-IV model. Standard errors for gaps between parents and differences across models are bootstrapped using 199 repetitions. The sample for the IV estimates includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were not required to be registered with a partner at the time (observations = 202,561; unique women = 11,666). The sample for the event study estimates includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were not required to be registered with a partner at the time, and eventually had at least one child (observations = 170,158; unique women = 9,726).

Table A9. Robustness: Comparison of long-run (age 11) fertility effects and child penalty estimates across models, without sample restriction that woman has partner

	LPR-IV (1)	Event (2)	Event-IV (3)	Event vs. Event-IV (2) - (3)
Gap	0.02	-0.18	-0.06	-0.12
	(0.21)	(0.05)	(0.10)	(0.10)
Women	0.05	-0.14	-0.02	-0.12
	(0.11)	(0.02)	(0.05)	(0.05)
Partners	0.03	0.04	0.04	0.00
	(0.21)	(0.05)	(0.10)	(0.10)

Note: Table shows estimates of earnings for women, partners, and the gap (women - partners), evaluated at $a = 11$ ($p = 11$ for LPR-IV). Column (1) shows estimates from the LPR-IV model, column (2) shows the estimates from the event-model, column (3) shows estimates from the event-IV model, and column (2) - (3) shows the difference between the event model and the event-IV model. Standard errors for gaps between parents and differences across models are bootstrapped using 199 repetitions. The sample for the IV estimates includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were not required to be registered with a partner at the time (observations = 202,561; unique women = 11,666). The sample for the event study estimates includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were not required to be registered with a partner at the time, and eventually had at least one child (observations = 170,158; unique women = 9,726).

A.8 Wage replacement

The Norwegian government provides substantial benefits to women during the latter part of pregnancy and the first year after birth. These benefits are meant to compensate for lost labor earnings and can be as high as 100 percent of lost earnings depending on labor market participation the year before birth. In order to give an estimate of the total earnings penalty carried by women having children we also show estimates when we replicate our baseline model using a broader measure of labor-related earnings and benefits that excludes capital gains and non-taxable transfers but includes sick leave and parental leave benefits. Figure A11 shows that while benefits substantially dampen the immediate effect of having children, the longer-run effect is very similar whether we include transfers or not.

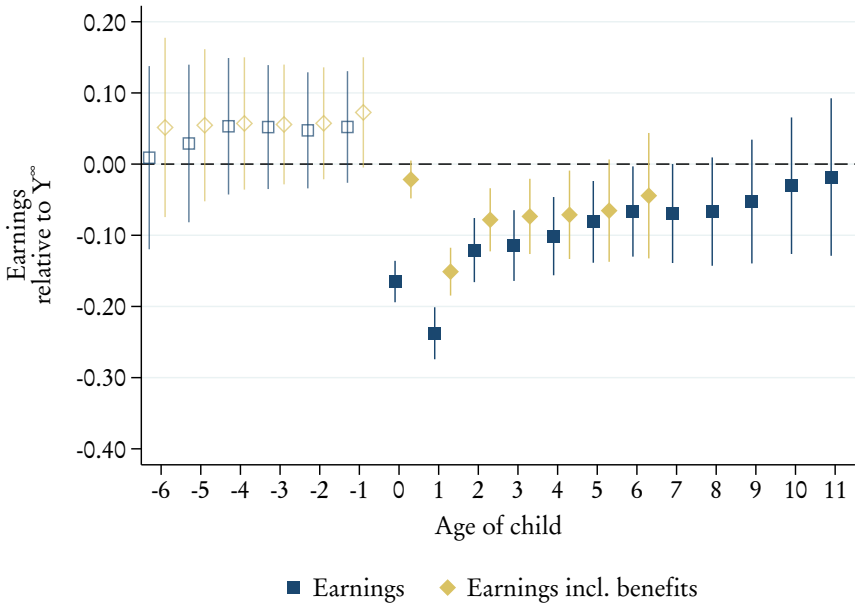


Figure A11. Earnings including benefits. Event-IV.

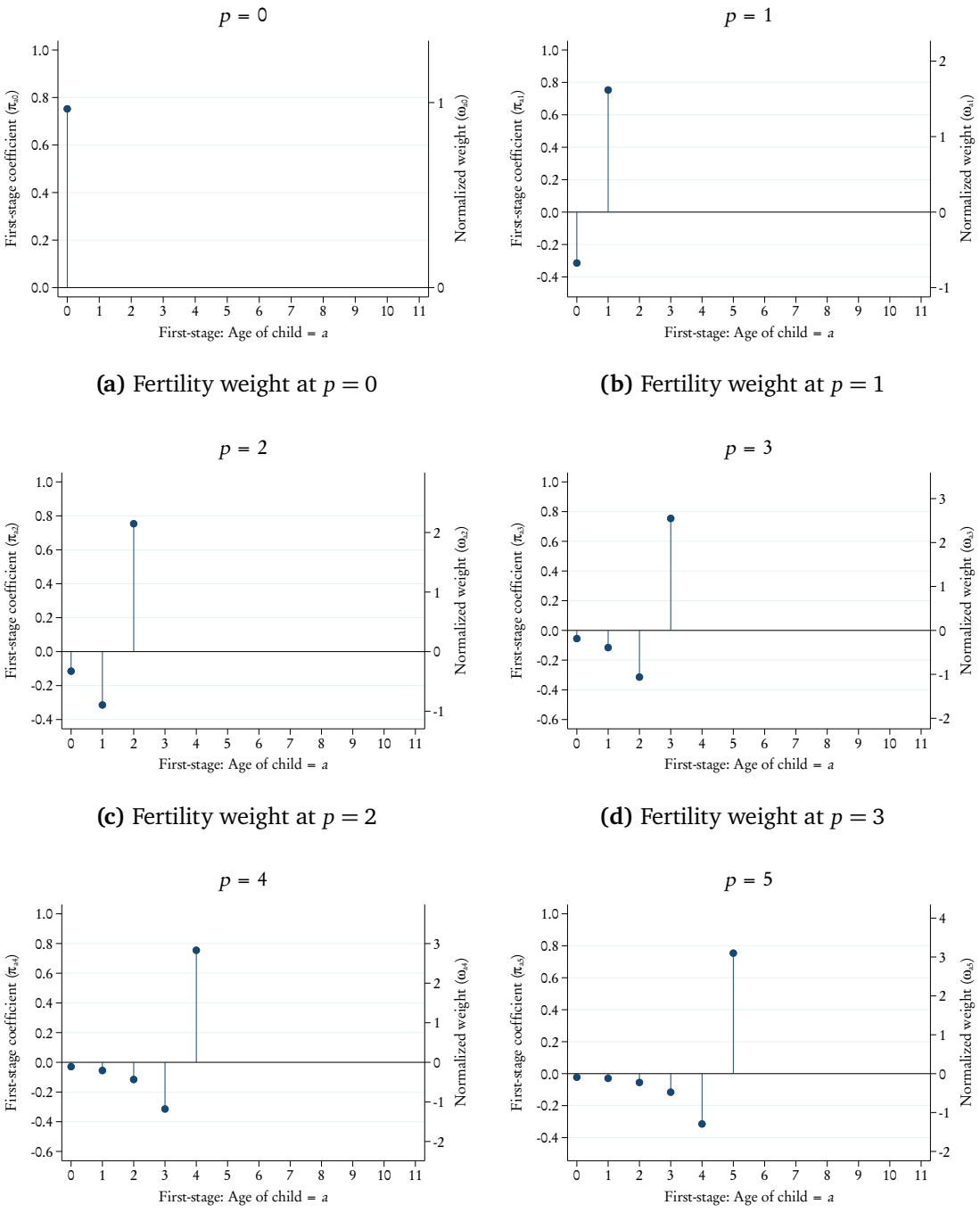
Note: Estimates from our Event-IV model as specified in equation (10, event-IV). Outcomes are earnings and earnings including benefits. Data on earnings including benefits are only available until 2017. Estimates are scaled relative to counterfactual earnings (Y^{∞}) as described in section 4.5. The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033)

A.9 Event-IV and LPR-IV

Table A10. First stage F-statistics for LPR-IV and event-IV.

	(1)	(2)
Years since IVF (column 1) / Age of child (column 2)	LPR-IV F-statistic	Event-IV F-statistic
-6		669
-5		788
-4		838
-3		865
-2		932
-1		955
0	800	972
1	807	960
2	807	951
3	811	959
4	811	953
5	809	951
6	757	830
7	695	751
8	615	628
9	533	524
10	447	417
11	349	284

Note: F-statistics for first-stages from the LPR-IV model (equation (5)) and the event-IV model (equation (11, FS, event-IV)). The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033).



(a) Fertility weight at $p = 0$

(b) Fertility weight at $p = 1$

(c) Fertility weight at $p = 2$

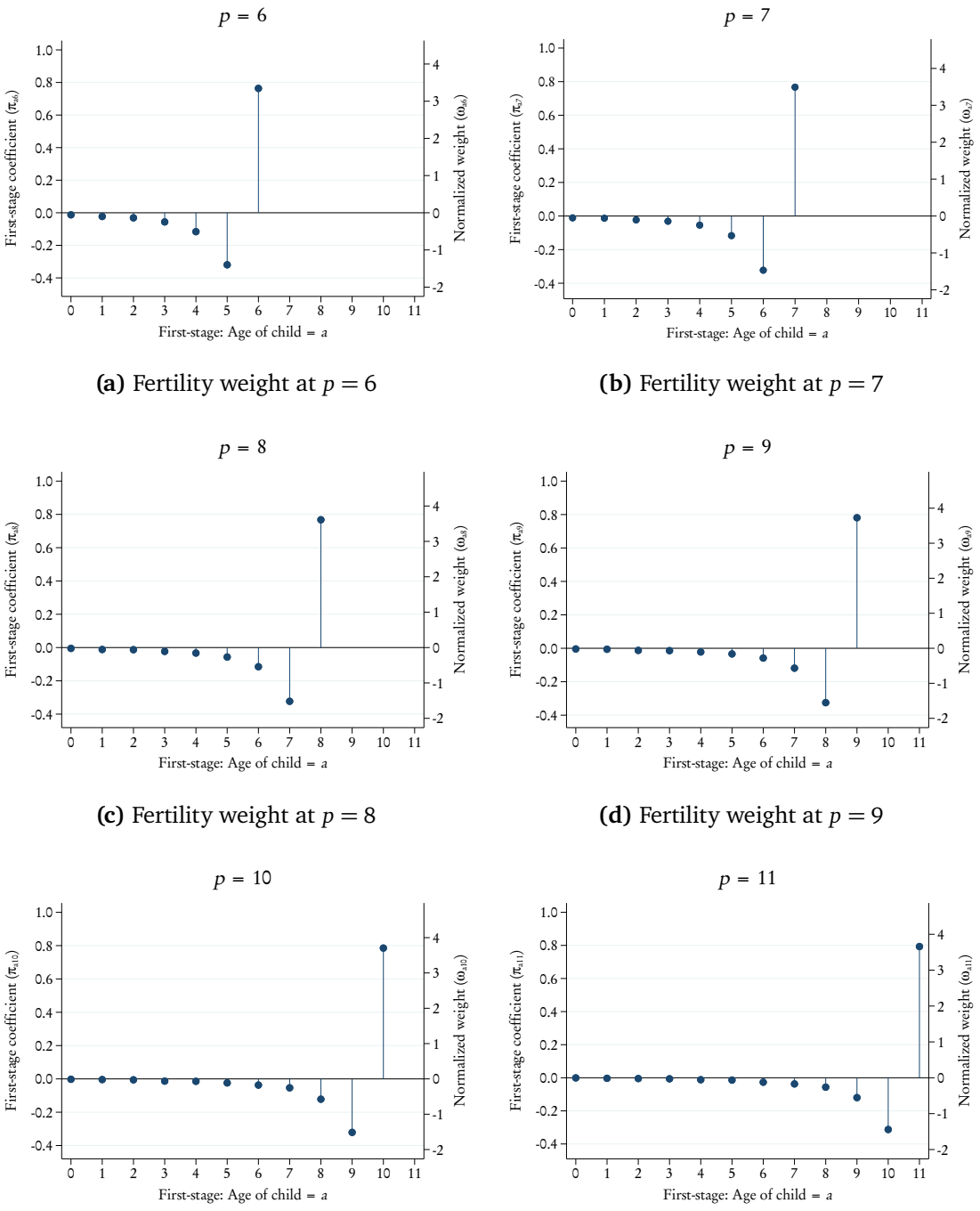
(d) Fertility weight at $p = 3$

(e) Fertility weight at $p = 4$

(f) Fertility weight at $p = 5$

Figure A12. Fertility weights

Note: Fertility weights as defined in Section 7.1. The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033).



(a) Fertility weight at $p = 6$

(b) Fertility weight at $p = 7$

(c) Fertility weight at $p = 8$

(d) Fertility weight at $p = 9$

(e) Fertility weight at $p = 10$

(f) Fertility weight at $p = 11$

Figure A13. Fertility weights

Note: Fertility weights as defined in Section 7.1. The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033).

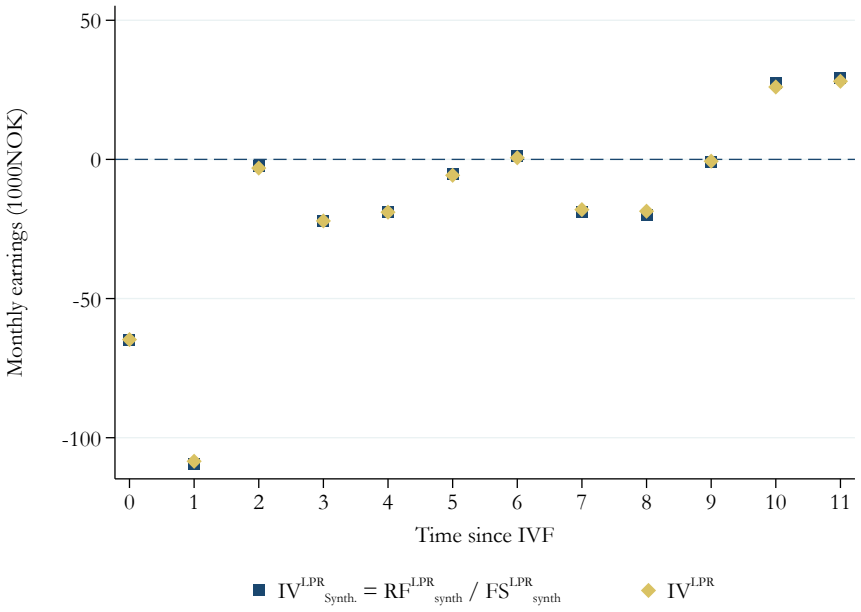


Figure A14. Combining results from LPR-IV and our event-IV model

Note: Figure shows results from our estimation of the IV model by Lundborg et al. (2017) alongside the event-IV estimates constructed from the reduced form and the first stages from our event-IV model in equation (10, event-IV) and (11, FS, event-IV). The sample is all women undergoing IVF treatment. The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033).

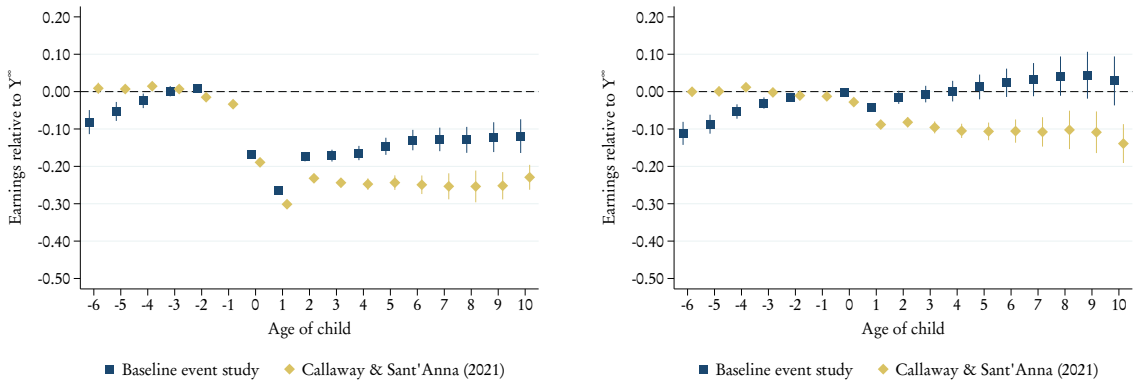
A.10 Complier characteristics

Table A11. Complier characteristics

	All IVF women			Compliers	
	Mean	Std.Dev.		Mean	Std.Dev.
Woman characteristics					
Age	33.99	(4.21)		34.11	(4.25)
Pre-IVF earnings	27.05	(17.00)		26.77	(16.97)
Education					
- Compulsory	0.14	(0.35)		0.15	(0.36)
- High School	0.24	(0.43)		0.25	(0.43)
- Bachelor	0.42	(0.49)		0.41	(0.49)
- Master	0.20	(0.40)		0.19	(0.39)
Sickness absence days	32.23	(72.68)	36.84	(76.27)	
GP visits	3.51	(3.87)	3.81	(4.02)	
Psychological symptoms	0.15	(0.36)	0.16	(0.37)	
Hospital days	6.80	(29.78)	8.11	(34.09)	
Partner characteristics					
Age	35.06	(6.10)		35.33	(6.23)
Education					
- Compulsory	0.17	(0.37)		0.17	(0.38)
- High School	0.39	(0.49)		0.39	(0.49)
- Bachelor	0.27	(0.45)		0.27	(0.44)
- Master	0.17	(0.38)		0.16	(0.37)
Earnings	36.96	(24.36)		36.58	(23.65)
(Estimated) number of women	10,033			7,527.08	

Note: Population and complier descriptive statistics evaluated one year after the first IVF trial. Complier mean and standard deviations computed using Abadie (2003) κ-weighting. The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; unique women = 10,033).

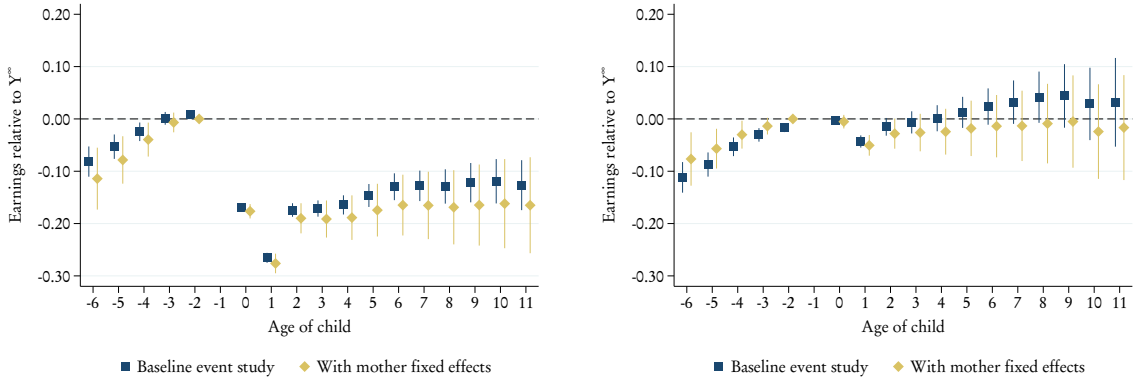
A.11 Alternative event-study estimators



(a) Woman's earnings **(b)** Partner's earnings

Figure A15. Results based on the treatment-cohort Callaway and Sant'Anna (2021) estimator

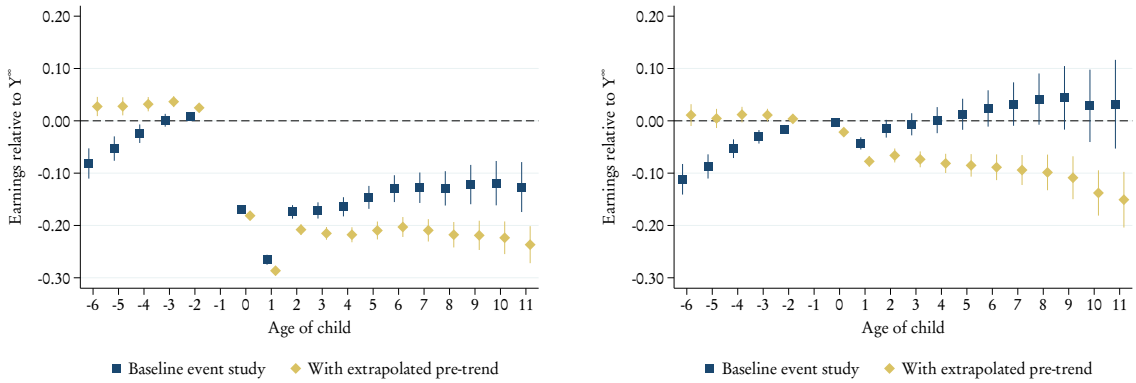
Note: This figure shows the estimated results from the event model using the conventional estimator as applied in f.e. Kleven et al. (2019) and results using the estimator proposed by Callaway and Sant'Anna (2021) with bootstrapped standard errors, allowing for effect heterogeneity by women's age at birth. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were registered with a partner at the time, and eventually had at least one child (observations = 145,571; unique women = 8,349).



(a) Woman's earnings **(b)** Partner's earnings

Figure A16. Results based on event-study specification with mother fixed-effects.

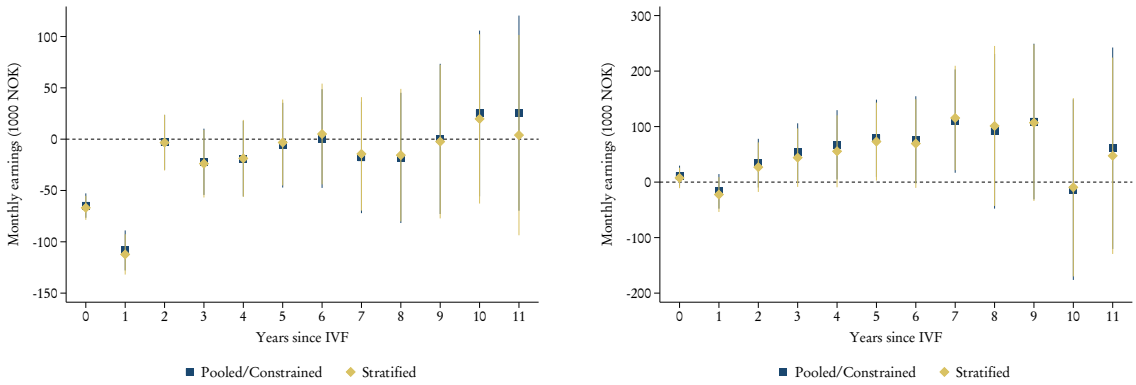
Note: OLS event study estimates from specification (2), but with mother fixed effects. Panel (a) shows effects separately for women and partners, panel (b) shows the difference between women and partners. Estimates are scaled relative to each gender's counterfactual earnings (Y^{∞}), as described in section 4.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were registered with a partner at the time, and eventually had at least one child (observations = 145,571; unique women = 8,349).



(a) Woman's earnings **(b)** Partner's earnings

Figure A17. Results based on Rambachan and Roth (2023) estimator

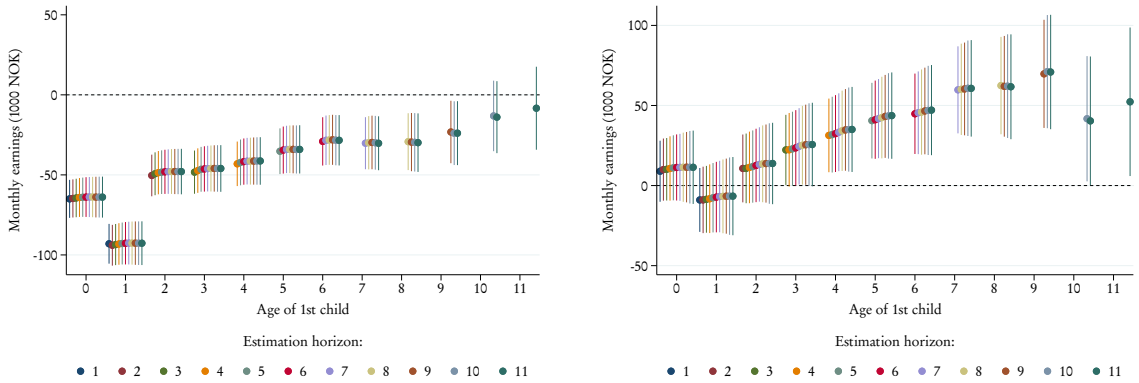
Note: This Figure A17 reports event-study estimates that adjust for the baseline of a linear extrapolation of the pre-trend into the post period following Rambachan and Roth (2023). Panel (a) shows effects separately for women and partners, panel (b) shows the difference between women and partners. Estimates are scaled relative to each gender's counterfactual earnings (Y^{∞}), as described in section 4.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were registered with a partner at the time, and eventually had at least one child (observations = 145,571; unique women = 8,349).



(a) Woman's earnings **(b)** Partner's earnings

Figure A18. Results by potential age of child

Note: Estimated effects of fertility on earnings using the LPR-IV model described in equation (4) on our data, separately by potential age of child (p). Panel (a) shows effects for women, panel (b) for partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in Section 4.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time.



(a) Woman's earnings

(b) Partner's earnings

Figure A19. Results by increasing horizon of potential ages p of child

Note: Estimated effects of age of child on earnings using the event-IV model described in equation (10, event-IV), estimated by increasing horizon of potential ages a. Panel (a) shows effects for women, panel (b) for partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 4.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time

Table A12. ICPC-2 Chapter P: Psychological Codes and Descriptions

Code	Description
Symptoms (P01–P29) and process codes (P30–P69)	
P01	Feeling anxious / nervous / tense
P02	Acute stress reaction
P03	Feeling depressed
P04	Feeling / behaving irritable / angry
P05	Senility, feeling / behaving old
P06	Sleep disturbance
P07	Sexual desire reduced
P08	Sexual fulfilment reduced
P09	Sexual preference concern
P10	Stammering / stuttering / tic
P11	Eating problem in child
P12	Bedwetting / enuresis
P13	Encopresis / bowel training problem
P15	Chronic alcohol misuse
P16	Acute alcohol misuse
P17	Tobacco abuse
P18	Medication abuse
P19	Drug / substance abuse
P20	Memory disturbance
P22	Child behaviour symptom / complaint
P23	Adolescent behaviour symptom / complaint
P24	Specific learning problem
P25	Phase of life problem, adult
P27	Fear of mental disorder
P28	Limited psychological function / disability
P29	Psychological symptom / complaint other
P30–P69	Process codes (consultations, tests, results, procedures)
Diagnoses and disorders (P70–P99)	
P70	Dementia / organic psychosis
P71	Organic psychosis, other
P72	Schizophrenia
P73	Affective psychosis (manic / depressive)
P74	Anxiety disorder / anxiety state
P75	Somatization disorder
P76	Depressive disorder / major depression
P77	Suicide / suicide attempt
P78	Neurasthenia / surmenage
P79	Phobia / obsessive-compulsive disorder
P80	Personality disorder
P81	Hyperkinetic disorder
P82	Post-traumatic stress disorder (PTSD)
P85	Mental retardation
P86	Eating disorder (anorexia / bulimia)
P98	Psychosis NOS / other
P99	Psychological disorder, other

Notes: The table lists all ICPC-2 codes under Chapter P (Psychological), grouped as in the official classification. Codes P01–P69 cover symptoms and process codes, while P70–P99 cover diagnoses and disorders. Our analysis uses one indicator including all Chapter P codes and another restricted to the subset of severe diagnoses and disorders.