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

Simon Bensnes† Ingrid Huitfeldt‡ Edwin Leuven§

February 7, 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 estimate a 15 percent child penalty, they differ in what drives this gap. The standard event study attributes the penalty 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). We thank Martin Andresen, Jon Fiva, James Heckman, Henrik Kleven, Magne Mogstad, Hessel Oosterbeek, and seminar participants at BI Norwegian Business School, University of Oslo, University of Chicago, Statistics Norway, OsloMet, 15th IZA and 2nd IZA/CREST conference on Labor market policy evaluation, and Zeuthen Workshop in Copenhagen 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 child birth, 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 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 event studies typically rely on exogeneity assumptions that allow for comparison of women who have children at different times.

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 requires the standard instrumental variable assumptions.

The event study and instrumental variable approaches differ not only in their underlying identifying assumptions, but they also recover 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) estimate the effect of having a child (of *any* age) at a given point in time since the IVF attempt. Their instrumental variable estimation setup (henceforth LPR-IV) leads not only to changes in the complier group over time as many women who fail a first IVF trial try again and are successful later but abstracts away from the fact that these women have children of different ages over time. 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. However, the size of the bias is not well understood, and it is not clear whether there is a direct mapping to the effect of having a child of a given age (the estimand of interest in the event study approach). We provide constructive results on these issues below which allows for

a reconciliation between the results from different identification approaches.

In our application we 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. We start the paper by formalizing and discussing the assumptions for the conventional event-study model and the LPR-IV model, before introducing an instrumental-variable based event-study model that combines these approaches (referred to as event-IV). The advantage of the event-study model relative to the LPR-IV model is the centering around birth, which addresses the potential violation of exclusion, and allows us to estimate dynamic fertility effects by the age of the child. Relative to the standard event-study model, the event-IV model allows us to address the potential omitted variables bias stemming from the endogeneity of fertility by exploiting information about the timing of the fertility attempt and the random variation generated by a successful IVF treatment in an instrumental variable setup.

We estimate and compare the earnings effects of fertility using the regular event-study, the LPR-IV, and the event-IV specifications. In 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 key policy implications rest on whether it is the mother 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.

When examining the separate estimates for women and partners, we find that while the event-study model suggests that nearly all of the child penalty is driven by women, the event-IV model finds that women account for less than half. More specifically, the event-study model indicates 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, suggesting minimal effects. The event-IV model falls in between, estimating a reduction of 7 percent. Turning to partners' earnings, the ordering of the estimates goes 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 mothers 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.

We explore the sources of bias and differences in estimates between models. We show that the estimated treatment effects in LPR-IV are latent mixtures of different dynamic (age-of-child specific) treatment effects and thus not directly comparable to event-study estimates. 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.

The difference between the standard event-study model and the event-IV estimates of the effect on mothers' long-run earnings is largely explained once we adjust earnings profiles for time since the IVF trial (a predetermined variable). Even without relying on instrumental variable assumptions, we therefore find that seemingly robust findings can substantially change when controlling for the timing of fertility attempts. 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. 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.

To understand how endogenous timing of fertility biases estimates from the standard event-study setup, we proceed by accounting for fertility timing when estimating the counterfactual earnings profiles. 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.

The type of selection we uncover not only means that event-study estimates can be biased even when pre-trends are parallel, but also that standard extrapolations of the pre-trend exacerbate the bias relative to the standard-event study specification. This goes against the common intuition that pre-trends are informative of violations of parallel trends in the treatment period (as for example formalized in Rambachan and Roth, 2023). Finally, we explore the role of confounding treatment effect heterogeneity as discussed in a series of recent advancements in the analysis of event study designs (see, e.g. Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; Borusyak et al., 2024; Goodman-Bacon, 2021; de Chaise-

martin and D'Haultfœuille, 2020). We implement the imputation estimators of Borusyak et al. (2024) and Callaway and Sant'Anna (2021) which both produce even more pronounced negative effects on maternal earnings than in the standard event study specification, even though they do not show evidence of any pretrends. These results are also consistent with the results based on the extrapolation of pre-trends following Rambachan and Roth (2023) which also exacerbate bias.[1]

While it is always challenging to extrapolate point estimates across different populations, we document that IVF women are observationally very similar to regular women of the same age and education level. Additionally, both standard event study fertility estimates and causal IV estimates suggest that the effects for IVF women and regular women are very similar. IVF births could however have different impacts on non-labor market outcomes such as divorce and mental health compared to regular births. In other contexts, instruments that rely on margins of eligibility may also induce disappointment effects on not receiving the treatment.[2] However, such effects affect only a small fraction of our sample, and a mediation analysis gives no indication that adjusting for these channels changes the estimates of children on labor market outcomes, nor does dropping all affected women from the estimation sample. Finally, recent evidence from Denmark which replicates part of our analysis and which can estimate fertility effects on the very long run (up to 24 years after birth) also confirm our findings (Lundborg et al., 2024).

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. 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. Gallen et al., 2023; Hotz et al., 2005; Cristia, 2008; Aguero and Marks, 2008; Miller, 2011). 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. Bronars and Grogger, 1994; Angrist and Evans, 1998; Rosenzweig and Wolpin, 1980).

In the next section we start by providing the relevant institutional background

[1] The estimator proposed by Callaway and Sant'Anna (2021) gives very similar results.

[2] Examples can be found in lottery or regression discontinuity based designs that study school assignment, housing assistance, job training or health care access.

information concerning IVF treatments as well as the social benefit system that will 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 different child penalty estimates are then reported and discussed in Section 6 after which Section 7 bridges and reconciles the different fertility effect estimates by documenting the sources of their differences and the nature of the 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.[3] 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 so-called single embryo transfer. This means that IVF had a low occurrence of multiple births (Bhalotra et al., 2019; Bhalotra and Clarke, 2019).

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.[4] 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 in-

[3] Anonymous donors are forbidden by law in Norway because every individual has a legal right to know the identity of their parents when turning 18 years old.

[4] This is broadly defined. The couple needs to be married or cohabiting in a marital-like relationship. Shared administrative registered address for 2 years can be used as documentation, as can cohabiting contracts. IVF treatment has been allowed for women with female partner since 2009.

serting 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 earnings/labor supply is mediated through both labor market mechanisms and social insurance/benefit systems. Norway has implemented comprehensive parental support systems since the 1970s (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.[5]

The support system extends beyond parental leave. Pregnant women could request welfare support if their working conditions posed potential risks to maternal or fetal health. Legal protections prohibited employers from pregnancy-based discrimination in hiring, promotion, and termination decisions. The system also provided 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—helped offset potential fertility-related earnings losses.

[5]The parental leave includes portions specifically designated for both mother and father. While eligibility depends on previous year's income, the criteria are relatively lenient, resulting in most parents qualifying.

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), exact time, date and place of admissions and discharges. We use these data to identify IVF trials from the 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 and reimbursement fees. From these data, we create a variable measuring the number of visits to the GP in a given year, as well as the subset of visits to the GP that are coded with a psychological symptom.[6] 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

[6]These visits fall under Chapter P "Psychological" in the ICPC-2 coding system. This chapter encompasses a range of mental health conditions and psychological issues commonly encountered in primary care, including mood disorders, anxiety, stress-related conditions, substance use disorders, and various behavioral and emotional problems.

hours.[7] We have access to these data for the period 2004 to 2022.[8]

We define four variables to capture individuals' labor market attachment. Our main outcome, *Earnings*, captures the yearly labor income.[9] *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.[10] *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 yearly earnings and report estimates for the other outcomes in the appendix.

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.[11] For comparison, we also construct a sample of mothers 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.

[7] Before 2015, the data on contracted hours are known to be of poor quality. We therefore assume that all workers in active employment spells work at least 4 hours per week. This affects reported hours for 0.07 percent of our sample. We also truncate very high hours (more than twice a standard full-time job, i.e. 162.5*2 hours per month) as these likely represent errors.

[8] A drawback of the employer-employee registry is that it does not cover income for self-employed or the benefits that are paid directly from the welfare office. This means that they do not fully reflect the insurance provided by the Norwegian benefit system. To investigate the role of such insurance, and the effect of fertility on disposable income given these relatively generous transfers, we additionally estimate earnings effects using the yearly tax files covering income from all sources.

[9] We adjust for inflation using 2015 as the base year.

[10] 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 and the magnitude of benefits such as old age pension, disability pension, and unemployment compensation. The basic amount is adjusted annually by the Norwegian Storting (parliament) to account for inflation and general wage growth.

[11] 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 mothers.

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

	(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)
Psychological symptoms	0.14	0.12	0.02	(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 mothers, 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.

Descriptive statistics are presented in Table 1.[12] Column (1) focuses on the sample of IVF women, while column (2) describes the non-IVF mothers. Labor market outcomes and health indicators are measured as an average over the four years prior to the first IVF trial or, for non-IVF mothers, prior to the approximate conception date (nine months before the birth).[13] Education is measured in the calendar year before the IVF attempt. Age is defined as the maternal age at the IVF attempt date.

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.[14] 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.[15]

The average age at first trial is just below 32, while non-IVF mothers have their first child at age 28. The education level is very similar but slightly higher for IVF 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 mothers earned 290,000 NOK per year. Among

[12] Our sample restrictions and restrictions on data availability create some attrition in the sample. We show the share of the women that can be observed in each period in Figure A1 in the appendix.

[13] The pre-period observation window for general practitioner and hospital visits are shorter for women who undergo their first treatment before 2010 and 2012, respectively, since data from GPs are available only since 2006, and hospital data since 2008.

[14] A pregnancy lasts for 38 weeks from conception (or 40 weeks from the first day of her last period), but we include the tenth month to ensure that we also retain women who go overdue.

[15] 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).

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 mothers, 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 mothers; and they on average 2.1 days per year with a visit to a hospital, compared to 1 day for non-IVF mothers. The average number of visits to the GP was about 2.5 per year for IVF women, and 1 for non-IVF mothers. There was only a small difference in the number of visits to the GP for a psychological symptoms; the average number of such visits was 0.14 per year for IVF-women and 0.2 for non-IVF women.

The average age of partners is 35 for IVF-women, compared to 31 for non-IVF mothers. 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 mothers earned 385,000 NOK and worked 0.78 FTEs per year.[16]

Compared to non-IVF mothers 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 mothers are more likely to have more than one child, their final fertility patters are overall quite similar.

4 Estimating the effects of fertility on labor market outcomes

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 birth, the event of interest. We can depart from the following general

[16] We track the earnings of the partner at the IVF attempt, regardless of whether the parents remain together or not.

potential outcomes

$$y_{it}^{\infty} = x_{it}'\phi + \tau_t + \epsilon_{it}^{\infty}$$
$$y_{it}^{a} = \delta_a + x_{it}'\phi + \tau_t + \epsilon_{it}^{\infty} + \epsilon_{it}^{a}$$

where superscript ∞ indicates the counterfactual of not (never) having a child, and a the counterfactual of having a child of age a. The controls x_{it} specify the counterfactual wage profile, where we control flexibly for mother's age using dummy variables, and calendar year dummies τ_t. The coefficients δ_a allow for age-of-child specific shifts in the outcome.

Observed outcomes map into potential outcomes as follows

$$\begin{aligned} y_{it} &= y_{it}^{\infty} + \sum_{a \geq 0} \mathbb{1}_{\{\text{age child}_{it}=a\}} (y_{it}^{a} - y_{it}^{\infty}) \\ &= \sum_{a \geq 0} \delta_a \mathbb{1}_{\{\text{age child}_{it}=a\}} + x_{it}'\phi + \tau_t + \epsilon_{it} \end{aligned} \tag{1}$$

where the child dummies $\mathbb{1}_{\{\text{age child}_{it}^{1}=a\}}$ equal one if the first child of woman i in calendar year t is a years old and are zero otherwise, and

$$\epsilon_{it} \equiv \epsilon_{it}^{\infty} + \sum_{a \geq 0} \mathbb{1}_{\{\text{age } \mathbf{1st} \text{ child}_{it}=a\}} \epsilon_{it}^{a}. \tag{2}$$

Equation (1) corresponds to a standard event-study specification.[17]

In practice the literature typically estimates equation (1) on samples of mothers while allowing for anticipation effects, and normalizes the counterfactual wage profile to a year prior to birth, $a = -1$, as in the following specification:

$$y_{it} = \sum_{a \neq -1} \delta_a \mathbb{1}_{\{\text{age child}_{it}=a\}} + x_{it}'\phi + \tau_t + \epsilon_{it} \tag{3}$$

where for notational convenience negative values of a refer to time before birth.[18] Our main outcome y_{it} and summary measure of women's labor supply is yearly earnings from work, but we consider additional outcomes in section A.5. In the full event-study specification, we report estimates from six years before birth up

[17] The standard event-study specification in the child penalty literature does not include individual or couple fixed-effects (e.g. Kleven et al., 2019). Some exceptions exist, e.g. Andresen and Nix, 2022, include couple fixed effects. We report event-study results from a model that includes couple fixed effects in Table A18 in the appendix. We also show results using the event-study estimator proposed by Borusyak et al. (2024), which includes individual fixed effects.

[18] This implies that the age-of-child dummies are formally defined as follows: $\mathbb{1}_{\{\text{age child}_{it}=a\}} \equiv \mathbb{1}_{\{\text{ child of mother } i \text{ born in year } t-a\}}$.

to eleven years after birth.

Assuming no heterogeneity in treatment effects, the counterfactual outcome profile in (3) is identified from the pre-birth wage profiles, and identification of δ_a, having children, is thus driven by differential timing of motherhood across women from the same cohort. The key assumption is therefore that fertility timing is exogenous conditional on the controls x_{it} and time-dummies τ_t. Consequently, if women with lower unobserved earnings potential tend to have children earlier than those with higher earnings potentials, the exogeneity assumption does not hold and the event-study overestimates the effect of having children.

Much of the child penalty literature examines the earnings gap between mothers and fathers rather than focusing on mothers' earnings alone. This approach has the advantage that it requires a weaker exogeneity assumption than examining maternal earnings in isolation. Rather than assuming that women's earnings trajectories would be identical regardless of when they have children, we only need to assume that the earnings gap between mothers and fathers would evolve similarly in the absence of children. However, this weaker assumption comes of course at a cost: it only allows one to examine the difference in outcomes between mothers and fathers, without being able to identify the separate impacts on women and men. We further discuss this distinction in Section 4.4.

4.2 Fertility effects of IVF (LPR-IV)

An alternative approach to identify fertility effects at the extensive margin that does not rely on the exogeneity assumption used in event studies comes from Lundborg et al. (2017) who argued that IVF can provide variation in fertility that is conditionally as good as random. They apply this in a two-stage least squares (2SLS) approach where fertility is instrumented with success in a first IVF trial. In the following, we refer to this model as LPR-IV.

Their outcome equation is as follows

$$y_{ip} = \gamma_p \text{Fertility}_{ip} + x'_{ip}\psi_p + \theta_p + u_{ip} \tag{4}$$

where time is now indexed by p which is the number of years since individual i's first IVF treatment. Consequently y_{ip} measures the outcome for woman i, but now observed p years after entering the IVF treatment. The explanatory variable of interest, Fertility_{ip}, equals one if woman i has a child p years after entering the IVF treatment and is zero otherwise. We follow Lundborg et al. (2017) and include

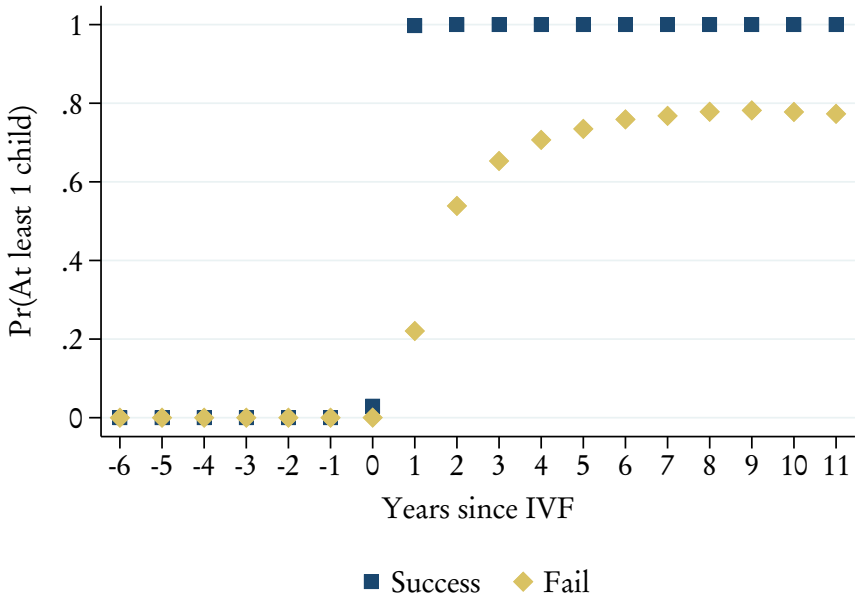


Figure 1. Fertility by success at first IVF trial

Note: Share of women having at least one child by year relative to first IVF treatment, grouped by success in first trial.

dummies for mother's age, x_{ip}, and calendar year. As we see below, the data are consistent with IVF success being exogenous conditional on mothers' age and education. In our implementation we therefore interact x_{ip} with the education level at the IVF trial to make sure that the conditional independence assumption underlying the instrument exogeneity condition is satisfied. In addition, equation (4) includes fixed-effects θ_p for years since woman i's IVF treatment.

Since fertility may correlate with unobserved determinants of the outcome, fertility is instrumented by the outcome of the IVF trial:

$$\text{Fertility}_{ip} = \pi_p \text{success}_i + x'_{ip}\lambda_p + \mu_p + w_{ip} \tag{5}$$

where the instrument success_i equals one if the IVF led to a birth. For IVF success to be a valid instrument, it should be as-good-as random conditional on x_{ip} and μ_p, and the outcome of the IVF trial can only affect the outcome through fertility (monotonicity is mechanically satisfied).

Figure 1 shows the fertility rates of women with a successful first IVF treatment and the fertility rates of women with a failed first IVF trial. For successful treatments fertility by definition jumps to 1. However, 87 percent of the women with a failure in the first IVF continue to a second attempt, and after a failure in the second IVF another 69 percent continues to a third IVF treatment. Both these

repeated IVF trials as well as non-IVF induced births lead to the catching up in fertility, and despite a failed first IVF, about 20 percent gives birth to a child one year later. After an additional two years this number has increased to 50 percent, and ultimately close to 76 percent of the women with a failed first IVF realizes motherhood.

In practice, many women therefore end up having children despite an unsuccessful first IVF trial. In instrumental-variable terminology this means that all women are compliers on the short-run (9 months) which implies that the first-stage coefficient π_p in (5) will be close to 1 for $p = 0$. The majority of women whose first IVF trials fail, try however again and ultimately conceive. They are therefore always-takers on the longer run and the share of compliers π_p drops as p increases. Lundborg et al. (2017) refer to this phenomenon as delayed fertility and point out that if the effect of children on earnings is larger when children are young then the fertility estimates γ_p will be a mixture of earnings loss and bias terms coming from delayed fertility. This is a violation of the exclusion restriction as IVF success not only affects fertility but also the age of the child. They also show that the fertility effects γ_p are likely to provide lower bounds on the underlying child penalties and can therefore still be informative about the impact of children on mothers' labor market outcomes. We show how this bias can be decomposed into event-IV child-penalty estimates in section 7.1.

While in theory it is possible to investigate subgroup heterogeneity with respect to the timing of subsequent IVF treatments and pregnancies, we follow Lundborg et al. (2017) and exploit only the first IVF. Using subsequent IVF trials as additional instruments does however not change our findings.[19]

4.3 Event-IV

The advantage of the standard event-study setup is that it recovers a well defined effect of having children, but it rests on the assumption of parallel trends conditional on observables. The advantage of the LPR-IV is that the variation in fertility is arguably more exogenous and transparent, but it recovers fertility effects that are weighted averages of the event study estimates, with unknown weights. We argue that combining these approaches has three distinct advantages.

First, centering time on the age of child as in the event-study setup rather than on time of the IVF trial carries the advantage that the treatment is well defined

[19]These results are available on request. Ketel et al. (2024) provide a framework for analyzing repeated treatment assignments in the LATE framework, and Ilciukas (2024) develops a non-parametric bounds approach with an application to IVF and labor supply.

and not a latent mixture of treatments arising from differential compliance over time (delayed fertility) and therefore addresses this potential violation of exclusion.

Second, and in a first step to address the concern that the timing of fertility is endogenous to labor supply, we note that IVF is also characterized by its timing. This allows us to control for whether a woman is "at risk" of giving birth. We therefore add the indicator variables $\mathbb{1}_{\{\text{time since IVF}_i=p\}}$ to equation (1):

$$y_{it} = \sum_{a\geq 0} \delta_a \mathbb{1}_{\{\text{age child}_{it}=a\}} + x_{it}'\phi + \tau_t + \sum_p \gamma_p \mathbb{1}_{\{\text{time since IVF}_i=p\}} + \epsilon_{it} \tag{6}$$

where x_{it} again contains dummies for mother's age and education.

Third, as documented above, about 20 percent of the IVF women realize fertility through other means than IVF alone. In a final step we therefore estimate (6) using 2SLS where we instrument $\mathbb{1}_{\{\text{age child}_{it}=a\}}$ with whether the woman was at risk of having an a-year-old through IVF and whether this attempt was successful:

$$\mathbb{1}_{\{\text{time since IVF}_i=p\}} \times \text{success}_i$$

and the resulting first stage is therefore as follows

$$\begin{aligned}\mathbb{1}_{\{\text{age child}_{it}=a\}} = &\sum_p \pi_{ap} \mathbb{1}_{\{\text{time since IVF}_i=p\}} \times \text{success}_i \\ &+ \sum_p \theta_{ap} \mathbb{1}_{\{\text{time since IVF}_i=p\}} + x_{it}'\tilde{\phi}_a + \tilde{\tau}_{at} + u_{iat}\end{aligned} \tag{7}$$

While the event-study specification of equation (3) is typically estimated on samples of women who eventually have children, we do not impose this restriction to our 2SLS sample as this would implicitly condition on IVF outcomes and violate instrument validity.

We estimate the child penalties in (6) relative to the counterfactual of not having a child ($a \geq 0$). This follows LPR and, as we will show below, allows for an exact mapping between their fertility effects and the dynamic effects of having a child of a given age. As a consequence we must estimate the pre-trends in the IV separately in the pre-IVF period (similar in spirit to Borusyak et al., 2024). The pretrend estimates are however, identical to those that would be estimated in a specification that anchors the estimates in $a = -1$. The child penalty estimates that are relative to $a \leq 0$ are in practice nearly identical to those that would be obtained relative to $a = -1$, because of the balance in the pre period resulting from the exogeneity of the instrument.

This event-IV specification combines the model based approach of the event study with the design based approach of LPR-IV. Section 7.1 below shows that the reduced form effects of IVF success can be interpreted as mixtures of child penalties due to delayed fertility. The identification of child penalties in the event IV approach rests on the assumption that child penalties are comparable for different complier groups. To see this note that in the first year following IVF the reduced form estimates the first-stage times δ_0, the effect of zero-year-olds for compliers. One year later the reduced form estimates a first-stage weighted average of the effect for one-year-olds for these women, δ_1, plus the effect of having a zero-year-old for women with delayed fertility. Because first-stage weights can be estimated, the one-year-old effect δ_1 is identified under the assumption that the zero-year-old effect is the same for women who are initially successful and for women with delayed fertility. This homogeneity assumption, which is also at work in the standard event study above, is technically an exclusion restriction.

To summarize, *i)* centering time on birth renders the treatment invariant to dynamic extensive margin fertility responses over time, *ii)* adjusting for timing accounts for the dynamic selection into the fertility attempt, and *iii)* the instrumentation addresses potential remaining unobserved variable bias due to other sources of fertility.[20]

4.4 Definitions of fertility effects and the child penalty

For ease of interpretation and comparability to the literature we focus on relative rather than absolute effects in cardinal units such as Norwegian Kroner. Following the literature, we scale the estimated effects relative to the average counterfactual outcome that would have been observed at the same point in time but in absence of the child. For women the estimand is therefore the following

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

Note that this means that a counterfactual outcome must be estimated for each age a of the child. In the standard OLS event study this is readily obtained by subtracting the fertility effects δ_a from the observed average wage of these women:

$$p_a^{women} = \frac{\delta_a^{women}}{E[y_{it} \mid \text{age child}_{it} = a, \text{women}] - \delta_a^{women}}$$

[20] 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.

For the IV approach we estimate the counterfactual outcome following Abadie (2003). The implementation with linear 2SLS involves re-estimating the fertility effects for each a where the outcome variable equals $-\mathbb{1}_{\{\text{age child}_{it} \neq a\}} y_{it}$.[21] If we denote the resulting counterfactual outcome for mothers by $\delta_a^{\infty,women}$ then the rescaled fertility effect IV equals

$$p_a^{women} = \frac{\delta_a^{women}}{\delta_a^{\infty,women}}$$

While our main focus is on the absolute impact of children on the labor market outcomes for mothers and their partners, the literature often focuses on the impact on the earnings difference between men and women, the so-called child penalty:

$$P_a = \delta_a^{women} - \delta_a^{men}$$

Estimating the difference $(\delta_a^{women} - \delta_a^{men})$ requires weaker identifying assumptions than estimating effects on maternal earnings (δ_a^{women}) alone, because as long as the estimates of δ_a^{women} and δ_a^{men} exhibit the *same* bias it will cancel out when taking the difference. More formally, if the child penalty is estimated with a bias which can be age-of-child (a) specific but the same for mothers and fathers:

$$\hat{\delta}_a^{parent} \xrightarrow{p} \delta_a^{parent} + Bias_a \text{ where } parent \in women, men$$

then the estimate of the difference is unbiased even if the estimates of the impact on levels are biased:

$$\hat{\delta}_a^{women} - \hat{\delta}_a^{men} \xrightarrow{p} \delta_a^{women} - \delta_a^{men}$$

This exogeneity assumption with respect to relative counterfactual earnings differences is typically referred to as a parallel-trend assumption.

In addition to our focus on the impact of children on the labor outcomes of mothers and partners, we also bring the empirical design outlined above to the estimation of the impact of children on earnings differences. We follow Kleven (2022) and focus on the age-specific difference in the scaled child penalty for mothers and fathers:

$$p_a = \frac{\delta_a^{women}}{\delta_a^{\infty,women}} - \frac{\delta_a^{men}}{\delta_a^{\infty,men}} \tag{8}$$

For this parameter of interest the event-study estimates rely on the assumption that if there is a bias, then it is a common *relative* bias in the fertility effects of

[21] To estimate the counterfactual outcome when having a child of age a requires changing the outcome variable to $\mathbb{1}_{\{\text{age child}_{it} = a\}} y_{it}$.

mothers and fathers:[22]

$$\frac{\hat{\delta}_a^{parent}}{\hat{\delta}_a^{\infty,parent}} \xrightarrow{p} \frac{\delta_a^{parent}}{\delta_a^{\infty,parent}} + Bias_a \text{ where } parent \in women, men \tag{9}$$

Finally, we use the Delta method to compute standard errors on the rescaled effects (c.f Appendix A.1).

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 demographics, labor market attachment and health.[23] As in the 2SLS specification in equation (6) and (7), 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. The regressions are estimated using averages from the four-year period preceding the first IVF trial for labor market and health measures.

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

[22] In contrast, Kleven et al. (2019) considered the estimated effect on the earnings difference scaled by the estimated counterfactual for the mother:

$$\hat{P}_a \equiv \frac{\hat{\delta}_a^{women} - \hat{\delta}_a^{men}}{\overline{y}_a - \hat{\delta}_a^{women}}$$

where $\overline{y}_a$ is the average income of women with a child of age a. Note that strictly speaking this targets not only a different estimand than Kleven (2022), but is also still biased under the parallel-trend assumption because the bias in $\hat{\delta}_a^{women}$ does not cancel out in the denominator.

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

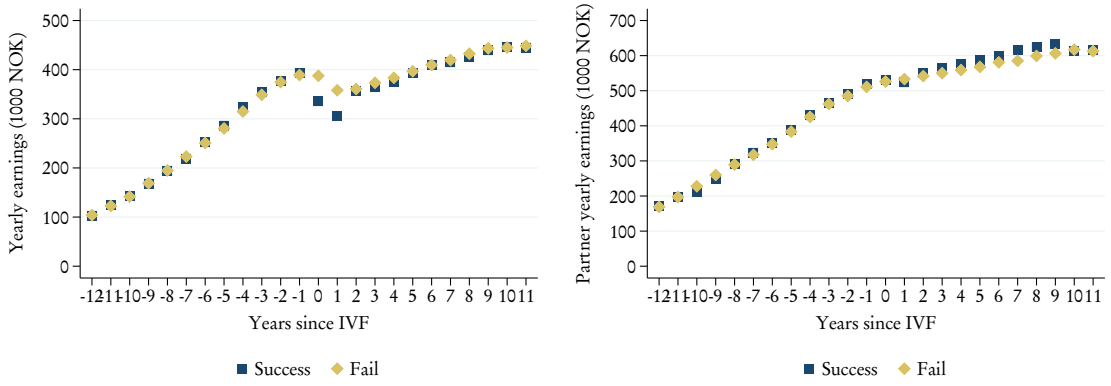
Table 2. Instrument validity

	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. 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 (6) and (7), both regressions include dummies for calendar time, time relative to IVF treatment, mother's age, and education. Joint Fs [p-value] refer to tests of joint significance of the characteristics shown in the table.

related 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 mother'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



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

Figure 2. Conditional independence of IVF success

Note: This figure plots average earnings for each year relative to the IVF trial. To rule out composition effects with respect to year, maternal education and age, the estimates are first completely stratified by year×age×education cells, before computing across-cell population level averages by IVF success and time since IVF.

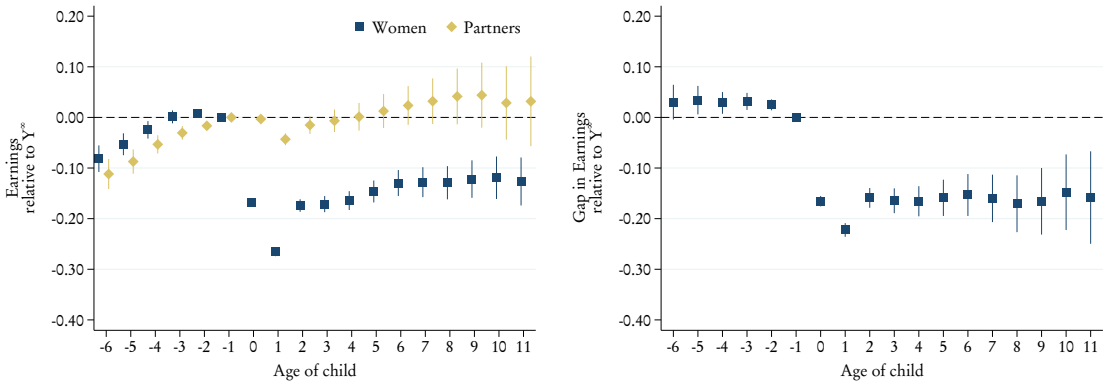
with concrete examples.

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 *trend*s, 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 does not affect any other variables than fertility directly. We discuss this, as well as issues related to 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



(a) Women and partners

(b) Gap

Figure 3. Earnings – Event study estimates

Note: OLS event study estimates from specification (3). 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 4.4. The samples consist of all mothers (i.e., women who eventually have at least one child) undergoing IVF treatment and their partners. Point estimates are presented in table form in Table A4.

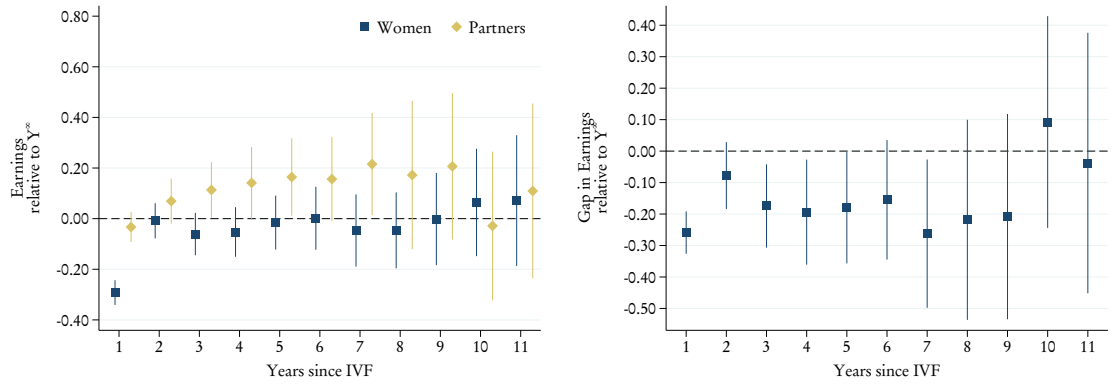
estimates of fertility since the IVF attempt (LPR-IV), and our specification that combines these two approaches (event-IV). For each model, we report estimates for mothers, partners, and the gap between the two (mothers minus partners, as in equation 8). Our discussion of long-term effects will focus 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 (3), estimated on IVF-mothers and their partners in Figure 3(a).[24] Both women and partners display a comparable pre-trend leading up to birth, indicating that women who have children earlier are on relatively steeper age-earnings profiles compared to those who have children later. Following birth, IVF mothers 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

[24]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 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 4.4. The samples consist of all women undergoing IVF treatment and their partners. Full set of point estimates are reported in Table A5 in the appendix.

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 A2 reports the estimated first stages from equation (4), 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 never below 500 in the first nine years

since IVF and are reported in appendix Table A8.

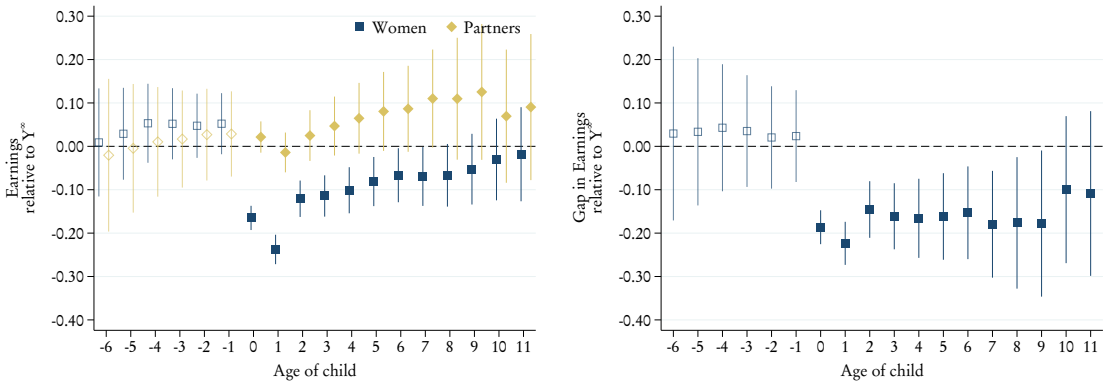
Figure 4(a) shows the IV estimates of equation (5), separately for women and their partners. Mother's earnings drop by about 30 percent in the year following the IVF treatment, but the effect quickly reverts to zero in the second year, at which level it remains for the remaining period. For comparison, Lundborg et al. (2017) find long run earnings losses for mothers at around 11 percent. As discussed in Section 4, this estimate is probably an upper bound (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 in 6 years. While the estimates are increasingly noisy they appear to be stable, estimating the average 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 mothers 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.

Not only do the event-study model and the LPR-IV model yield different effects when looking at earnings gaps between partners, the point estimates for mothers and their partners are also strikingly different. Where the event-study finds that mother's earnings fall in the neighborhood of 13 percent, the LPR-IV specification shows that penalties 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 7. F-statistics for the first stages are reported in appendix Table A8 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



(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 (7). Panel (a) shows effects separately for mothers and partners, panel (b) shows the difference between mothers and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 4.4. The samples consist of all women undergoing IVF treatment and their partners. Full set of estimates in table form are reported in Table A6 in the appendix.

run. Figure 5(b) plots the estimated gap between mothers 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 A7 and A8 report results for additional labor market outcomes (hours, employment and hourly wages) for the event-IV model.[25] 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 A8, 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 A8 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

[25] Results for the same outcomes for the other models are reported in Figures A3 - A6

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 A12 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 mother, 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 report analogous results for age eleven ($a = 11$) in appendix Table A7. 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 across the event models. 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 mothers 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

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
	(0.10)	(0.02)	(0.05)	(0.05)
Mother	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 mother, partner, and the gap (mother - partner), evaluated at $p = 6$. 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.

driven by mothers, 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 mothers and partners, the event-IV model paints a different picture than the event-study model. For mothers, it estimates only a small long-run 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 mothers 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 mothers 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

Our event-IV estimates can be mapped into the results from the LPR-IV. While the event-IV model estimates fertility effects by the age of the child, and the LPR-IV model by time since the IVF treatment (the "potential age of child"), the instruments and outcomes are identical. This implies that the reduced forms are identical. This means that we should be able to map our first-stage and event-IV estimates, which are centered by the age of the child, into fertility effects that are centered on time relative to IVF.

We do this by noticing that fertility is defined by the following identity

$$\text{Fertility}_{ip} \equiv \sum_{a \geq 0} \mathbb{1}_{\{\text{time since IVF}_i = p\}} \mathbb{1}_{\{\text{age } \mathbf{1st} \text{ child}_{it} = a\}} \tag{10}$$

Substituting the event-IV first-stages (7) into (10) we get

$$\begin{aligned}\text{Fertility}_{ip} &= \sum_{a\geq 0} \mathbb{1}_{\{\text{time since IVF}_i=p\}} \Big(\sum_l \pi_{al} \mathbb{1}_{\{\text{time since IVF}_i=l\}} \times \text{success}_i\Big) \\ &= \left(\sum_{a\geq 0} \pi_{ap}\right) \mathbb{1}_{\{\text{time since IVF}_i=p\}} \times \text{success}_i\end{aligned}$$

where the second line follows from the fact that all interactions cancel except when $p = l$. This expression shows that there is a one-to-one mapping between the first-stage coefficients of LPR-IV who condition on time since IVF and the event-study first-stage coefficients:

$$\text{Fertility}_{ip} = \pi_p \text{success}_i$$

where

$$\pi_p = \sum_{a\geq 0} \pi_{ap}$$

We can similarly derive the reduced form of the event-IV setup as follows

$$\begin{aligned}y_{it} &= \sum_{a\geq 0} \delta_a \mathbb{1}_{\{\text{age 1st child}_{it}=a\}} + \dots \\ &= \sum_{a\geq 0} \delta_a \Big(\sum_p \pi_{ap} \mathbb{1}_{\{\text{time since IVF}_i=p\}} \times \text{success}_i\Big) + \dots \\ &= \sum_p \mathbb{1}_{\{\text{time since IVF}_i=p\}} \sum_{a\geq 0} \delta_a \pi_{ap} \text{success}_i + \dots\end{aligned}$$

which shows that the reduced form coefficient of LPR-IV p years after IVF equals $\sum_{a\geq 0} \delta_a \pi_{ap}$, and that their IV estimate of fertility p years after IVF which is the ratio of the reduced form and first-stage coefficient can be written as

$$\gamma_p = \frac{\sum_{a\geq 0} \pi_{ap} \delta_a}{\sum_{a\geq 0} \pi_{ap}} = \sum_{a\geq 0} \omega_{ap} \delta_a \tag{11}$$

This shows that the effect of having children γ_p is a weighted average of the event study estimates δ_a where the weights are the normalized first stage coefficients $\omega_{ap} \equiv \pi_{ap} / \sum_{a\geq 0} \pi_{ap}$. While the weight on $\delta_{a=p}$, the effect of having a p-year-old p years after the IVF attempt is positive, we find that the weights on the estimates for younger children ($\delta_{a<p}$) are negative. The intuition is that mothers with delayed fertility have younger children.

We report the estimated weights in (11) 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 at year 6 after

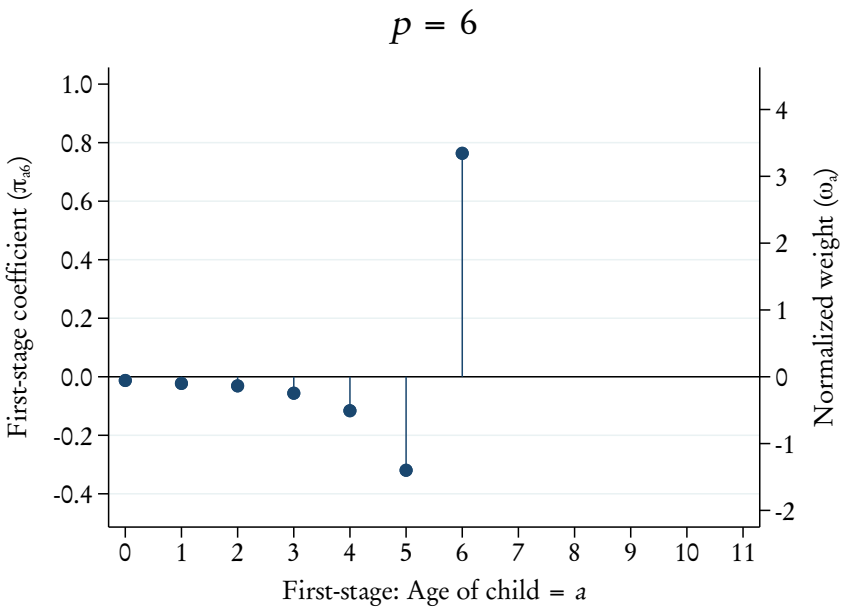


Figure 6. Mapping the first stages of LPR-IV to event-IV

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 potental 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 potental age $p = 6$).

IVF (π_{a5}). The right-hand y-axis shows the normalized weight for each first-stage (ω_{a5}). 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 A13 and A14. 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 A15 where we plot the results for mother's earnings from the LPR-IV model along with the rescaled estimates constructed from the reduced form and the rescaled

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 fertility into dynamic treatment effects of having a child of a specific age.[26]

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 mothers (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 (3).

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 mothers 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

[26] Note that the equivalence requires that all controls are interacted with $\mathbb{1}_{\{\text{time since IVF}_i = p\}}$. Our 2SLS event-study specification in (6) is separable and thus more parsimonious, which explains why the estimate do not exactly line up, but also shows that this has not consequences for our estimates.

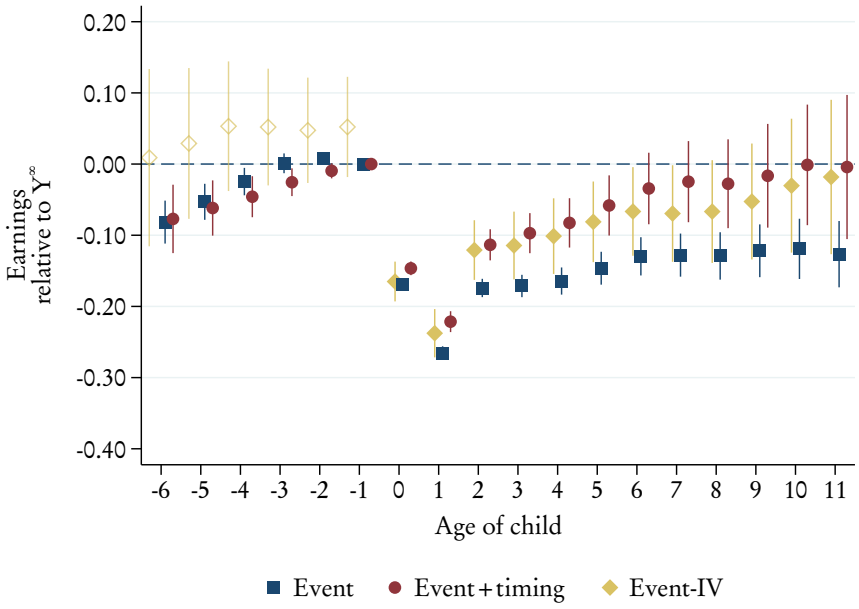


Figure 7. Event vs. Event-IV estimates

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 4.4.

children.

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-

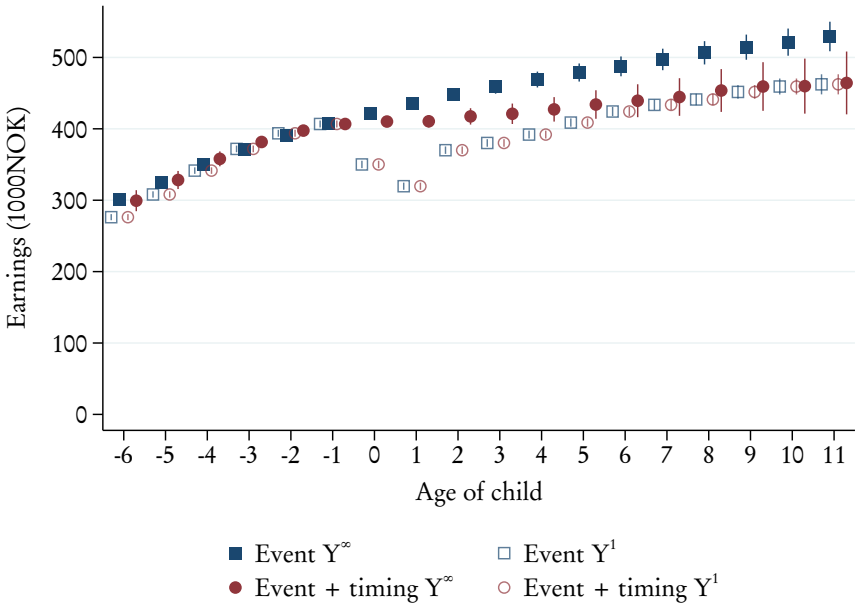


Figure 8. Counterfactual earnings profiles – Event-study estimates.

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.

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 A9 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 the idea that pre-trends are informative about violations of parallel trends and propose checks to assess how sensitive results are deviations from the pre-trends after treatment. Appendix Figure A16 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 (3), 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, we apply the estimator Callaway and Sant'Anna (2021) to our sample of IVF women, and the imputation approach of Borusyak et al. (2024). The two approaches differ in that the Borusyak et al. (2024) approach estimates the counterfactual Y^{∞} on all the not-yet-treated observations and includes individual fixed effects which subsume confounding level effects associated with fertility timing. The Callaway and Sant'Anna (2021) 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).[27] In both approaches we allow for heterogeneity by mother's age at birth (rather than calendar year).

Results are plotted in appendix Figures A16 and A17, along with the conventional event study estimates. Figure A17 shows that for both mothers and their partners the Callaway and Sant'Anna (2021) approach lines up the pre-trends, and estimates somewhat larger negative effects of children on earnings. The Borusyak et al. (2024) estimates shown in Figure A17 lead to even more negative estimates.[28]

The differences between these two estimators are best understood from the functional form assumptions they impose on the counterfactual. The Callaway and Sant'Anna (2021) approach is relatively non-parametric, while Borusyak et al. (2024) approach – like the standard event study estimator – relies more on the specific functional form assumed in the imputation step in the non-treated sample. In this context, both approaches aggravate the bias relative to the standard event study specification. This is consistent with the results based on the extrapolation of pre-trends above. It also highlights 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.

8 External validity

The analysis in this paper and the insights it provides depend on two key characteristics that are uniquely provided by the context of IVF, namely i) information about the timing of fertility and ii) conditional randomization of birth. To put our findings in perspective, this section compares the context of our study to that of the broader population. This raises two questions: the first one concerns comparability of population, and thus asks whether IVF women are different from regular women. The second is about context and asks whether IVF births are different from regular births. We discuss these in turn.

8.1 Are IVF women different from regular women?

In terms of external validity, one might wonder if there are differences between IVF women and non-IVF women that could lead to different responses to having

[27] We use Stata's implementation of Callaway and Sant'Anna (2021), and the Stata package provided by Borusyak et al. (2024).

[28] Estimating the standard event study with mother fixed gives qualitatively similar results (see Figure A18).

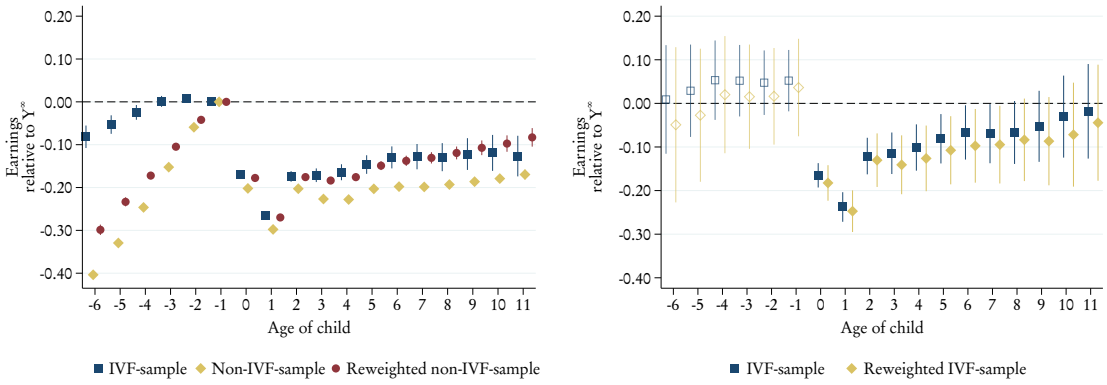
children. Such differences could both occur because the selection into timing of fertility is different with respect to potential labor market outcomes, or because having children induces different labor market outcomes.

We start by considering how the timing of fertility varies between samples. In Table 1 we saw that women undergoing IVF treatments tend to be older and more educated than mothers in the non-IVF sample. IVF women also end up having somewhat fewer children than women in the broader population. Fertility differences at the intensive margin are, however, completely explained by differences in age and education. In Table A2 we show that when we reweight the non-IVF sample to match the IVF sample in age and education at the time of conception they end up with virtually the same number of children as IVF women who have at least one child.[29] Table A3 reports other descriptive statistics for the reweighted non-IVF sample. Conditional on age and education, there are some remaining differences: IVF women and their partners have somewhat higher earnings and IVF women have a bit more sickness absence. IVF women are therefore somewhat more positively selected, but overall the differences appear to be relatively minor.

An alternative approach to investigate the importance of across-population heterogeneity would be to compare standard event study estimates for IVF women to those for regular women. Differential selection or anticipation would commonly be diagnosed from inspecting pre-trends.

Event study estimates for mothers are reported for both populations in Figure 9a. The first thing to note are the steeper pre-trends for the regular population which suggest stronger selection. Evidence in support of this comes from the Norwegian Mother and Child Cohort Study (Magnus et al., 2006). These data indicate that 82 percent of mothers in the general population in Norway report that their child resulted from a planned pregnancy. This implies that while IVF women and non-IVF women are relatively similar in that they are planning their pregnancies, about one out of five pregnancies in the broader population will be non-planned. The steeper pre-trend in the regular population is consistent with negative selection into pregnancy for women who have their unplanned children at earlier ages. This interpretation is also in line with the results of Gallen et al. (2023) who use Swedish data and find that unplanned pregnancies have larger negative effects for younger women. Finally, while women undergoing IVF are actively planning their fertility, they are likely more constrained by external fac-

[29] Specifically, we weigh the observations by the inverse propensity score, with propensity scores estimated from a probit model using a fully saturated model with age and education variables.



(a) Event-study specification

(b) Event-IV specification

Figure 9. Earnings. Reweighted estimates

Note: Panel (a) shows event-study estimates for the IVF-sample, non-IVF-sample, and non-IVF sample reweighted to match the composition of the IVF-sample. Effects are estimated in samples of mothers only, i.e., women who eventually have at least one child. Panel (b) shows event-IV estimates for the IVF-sample compared to event-IV estimates reweighted to match the composition of the non-IVF sample. Effects are estimated in the sample of all women undergoing IVF treatment. Weights are inverse propensity scores, estimated using a probit model that is fully saturated with age and education. Estimates are scaled relative to each gender's counterfactual earnings (Y^∞), as described in section 4.4.

tors regarding timing compared to women who conceive naturally. This reduced flexibility may result in less scope for selective choices, potentially explaining the greater negative selection observed in the regular population. Taken together these factors suggest that timing-bias is also a potential concern for the broader population.

A comparison of the post-birth fertility effects can also shed light on heterogeneity across populations. Figure 9a shows that the non-IVF mothers experience a larger drop in earnings following birth than the IVF mothers. However, reweighting the non-IVF sample makes the event 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 having children (or bias) is very different from women in 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 mothers 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 heterogene-

ity across these traits would not lead to dramatically different estimates for the overall population.

We interpret these results as indicating that effect heterogeneity in (observable) differences is unlikely to drive the differences between the event-IV estimates for IVF women and the event estimates for non-IVF women. Given the similarity of event-study estimates across groups and our finding that removal of selection bias reduces the earnings effects for women in the IVF-sample, it seems likely that event estimates for non-IVF women are also overestimated (i.e. too negative). However, without an instrument, it is hard to say anything about the magnitude of this bias.

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

A comparison of populations based on treatment effect estimates also relies on the treatment being the same. The second comparability issue therefore concerns the question of whether IVF births and the resulting impacts on parents are somehow different from regular births.

The IVF literature that is interested in the effects of fertility uses instrumental-variable estimators. In general, instrumental variables must be exogenous, relevant, induce monotone responses, and satisfy exclusion restrictions. Conditional exogeneity was documented above, and relevance and monotonicity mechanically hold. Exclusion was also discussed above in the context of delayed fertility.

A potential exclusion-like concern is "disappointment" following a failed IVF trial. A "disappointment effect" of the instrument is not unique to the IVF setting. For example, in Gallen et al. (2023) who estimate the effect of unplanned children resulting of failed contraception, "disappointment" may ensue when having the child rather than when not conceiving as with IVF. In other contexts, instruments that rely on margins of eligibility may also induce disappointment effects when not receiving the treatment. Examples can be found in lottery or regression discontinuity based designs that study school assignment, housing assistance, job training or health care access.

However, given that with IVF the instrument is actually having a child, any impact is defined relative to birth and is as such by definition a fertility effect. We therefore think of this issue not so much as a violation of exclusion, but rather as one of the many potential outcomes influenced by fertility. For example, in addition to affecting labor supply, parents may re-optimize their time spent at home and work and adjust to life with a child. In the context of this study, the question of whether IVF births differ from regular births pertains to whether the mediating

effects of fertility through other outcomes significantly impact the labor market responses reported above.

To investigate potential differences between IVF and natural births, we examine two key non-labor market outcomes: psychological well-being and marital stability. While some early studies suggested that failed IVF attempts could adversely affect mental health (e.g. Verhaak et al., 2005), a growing body of evidence finds minimal or no effects on various mental health measures, including self-reported well-being, medication use, and clinical depression (e.g. Verhaak et al., 2007; Agerbo et al., 2013; Baldur-Felskov et al., 2013; Pedro et al., 2019; Yli-Kuha et al., 2010; Lundborg et al., 2024).

Our analysis confirms these findings. Figure A9 tracks visits to general practitioners for psychological symptoms, including mood disorders, anxiety, stress-related conditions, substance use disorders, and behavioral or emotional problems. While there is indeed an impact on mental well-being, the effect is short-lived and after three years we find no systematic differences between women with and without children. For divorce outcomes (figureA9a), we find that having children reduces the probability of divorce by about 5 percentage points in the long run, though this effect follows a different temporal pattern than the psychological effects.

A formal mediation analysis suggests that these other outcomes do not substantially impact the main conclusions. As shown in Figure A10, controlling for both divorce and mental health leaves our estimated labor market impacts virtually unchanged. Moreover, these effects appear in only a small fraction of our sample. In Figure A11, we show that when we estimate our model in the IVF sample after excluding all observations with a psychological diagnosis, and then after excluding all women with a psychological diagnosis post-IVF treatment, the effects of children remain virtually identical to the baseline model. Taken together we interpret these findings as suggesting that while IVF births differ from natural births in some dimensions, they are unlikely to meaningfully affect our core findings about labor market responses.

8.3 Summary external validity

While it is always hard to extrapolate point estimates with confidence across populations, similarity in population and context is often used to argue for a degree of external validity. We documented above that while IVF women are somewhat positively selected compared to regular women, they are observationally very similar to regular women of similar age and education.

Regular event study fertility estimates and causal IV estimates also suggest that effects for IVF and regular women are comparable. Pre-trends diverge however, but conventional interpretation of pre-trends would lead one to conclude that the regular event study estimates for regular women suffer from more negative selection bias than those for IVF women. This would imply that the estimates for IVF women can be interpreted as lower bounds on the effects for the broader population.

IVF births could affect parents differently than regular births. This is often raised as a concern for non-labor market outcomes. Women who try to conceive without IVF and fail may also experience this as a setback. Similarly, regular couples who are struggling to conceive may also experience higher risk of divorce. To what extent IVF births are different from regular births in this respect is an empirical question, which is impossible to answer without better data and information on fertility plans for a representative sample. There is however no indication that these channels substantially matter for the effects of labor market outcomes which are the focus of this study.

Finally, there is the question whether our results extrapolate to other contexts than Norway. Recent evidence on this comes from Lundborg et al. (2024) who estimate fertility effects on the very long run for Denmark using their original LPR-IV specification. They also find that effects for mothers go to zero over time. Finally they also replicate our findings for the standard event study above which also exhibits substantial bias on the long run.

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 mothers 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 mothers 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 mothers 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.

Agerbo, E., P. B. Mortensen, and T. Munk-Olsen (2013). Childlessness, parental mortality and psychiatric illness: a natural experiment based on in vitro fertility treatment and adoption. *J Epidemiol Community Health 67*(4), 374–376.

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.

Baldur-Felskov, B., S. K. Kjaer, V. Albieri, M. Steding-Jessen, T. Kjaer, C. Johansen, S. O. Dalton, and A. Jensen (2013). Psychiatric disorders in women with fertility problems: results from a large danish register-based cohort study. *Human reproduction 28*(3), 683–690.

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.

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.

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

Ilciukas, J. (2024). Parenthood timing and gender inequality. Unpublished working paper, University of Amsterdam.

Ketel, N., E. Leuven, H. Oosterbeek, and B. Van der Klaauw (2024). Using principal stratification to analyze repeated treatment assignment. Unpublished working paper, University of Oslo.

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.

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.

Pedro, J., D. Vassard, G. M. H. Malling, C. Ø. Hougaard, L. Schmidt, and M. V. Martins (2019). Infertility-related stress and the risk of antidepressants prescription in women: a 10-year register study. *Human Reproduction 34*(8), 1505–1513.

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.

Verhaak, C. M., J. M. Smeenk, A. W. Evers, J. A. Kremer, F. W. Kraaimaat, and D. D. Braat (2007). Women's emotional adjustment to ivf: a systematic review of 25 years of research. *Human reproduction update 13*(1), 27–36.

Verhaak, C. M., J. M. Smeenk, A. Van Minnen, J. A. Kremer, and F. W. Kraaimaat (2005). A longitudinal, prospective study on emotional adjustment before, during and after consecutive fertility treatment cycles. *Human reproduction 20*(8), 2253–2260.

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

Yli-Kuha, A.-N., M. Gissler, R. Klemetti, R. Luoto, E. Koivisto, and E. Hemminki (2010). Psychiatric disorders leading to hospitalization before and after infertility treatments. *Human Reproduction 25*(8), 2018–2023.

A Appendix For Online Publication

A.1 Standard errors on rescaled estimates

Denote the rescaled estimate by x:

$$x = \frac{y^1 - y^0}{y^0} \equiv \frac{\delta}{y^0}$$

The Delta method gives

$$V(x) = \begin{pmatrix} \partial x/\partial \delta \\ \partial x/\partial y^0 \end{pmatrix}' V \begin{pmatrix} \delta \\ y^0 \end{pmatrix} \begin{pmatrix} \partial x/\partial \delta \\ \partial x/\partial y^0 \end{pmatrix}$$

where

$$V \begin{pmatrix} \delta \\ y^0 \end{pmatrix} = \begin{pmatrix} V(\delta) & cov(\delta, y^0) \\ & V(y^0) \end{pmatrix} = \begin{pmatrix} V(\delta) & (V(y^1) - V(y^0) - V(\delta))/2 \\ & V(y^0) \end{pmatrix}$$

since

$$\begin{aligned} V(\delta) &= V(y^1) + V(y^0) - 2cov(y^1, y^0) \\ \Rightarrow cov(y^1, y^0) &= (V(y^1) + V(y^0) - V(\delta))/2 \\ &\text{from this we get} \\ cov(\delta, y^0) &= cov(y^1, y^0) - V(y^0) \\ &= (V(y^1) - V(y^0) - V(\delta))/2 \end{aligned}$$

we also have that

$$\begin{pmatrix} \partial x/\partial \delta \\ \partial x/\partial y^0 \end{pmatrix} = \begin{pmatrix} 1/y^0 \\ -x/y^0 \end{pmatrix}$$

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

$$\begin{aligned} V(x) &= (V(\delta) - 2 \cdot x \cdot cov(\delta, y^0) + x^2 V(y^0))/(y^0)^2 \\ &= (V(\delta) - x \cdot (V(y^1) - V(y^0) - V(\delta)) + x^2 V(y^0))/(y^0)^2 \end{aligned}$$

where $V(\delta), V(y^1)$ and $V(y^0)$, all come from separate 2SLS regressions as outlined in section 4.3.

A.2 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 women who had at least one IVF trial over the period 2009 to 2016, 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 mothers	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 mothers, prior to the approximate conception date. Age and education are measured the year before the IVF treatment.

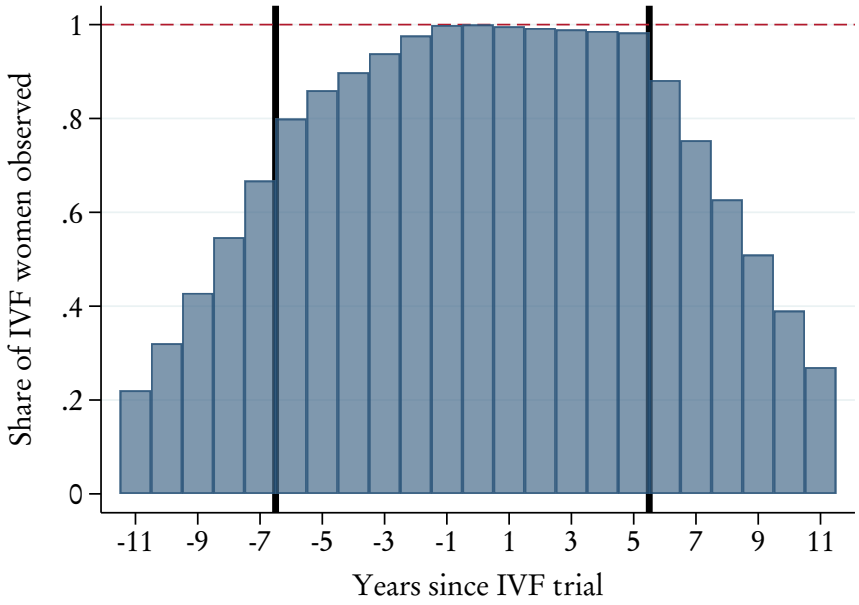


Figure A1. Distribution of observations over time since IVF

Note: Bars represent the share of observations in each year relative to the IVF year. All women are observed in relative year 0. In the years between the vertical lines, missing observations are due to women not being observed due to residency or age requirements. Observations outside the vertical lines can be missing either due to age and residency requirements, but also mechanically from data not being available for all calendar years.

A.3 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.

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.

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.

A.4 LPR-IV First Stage

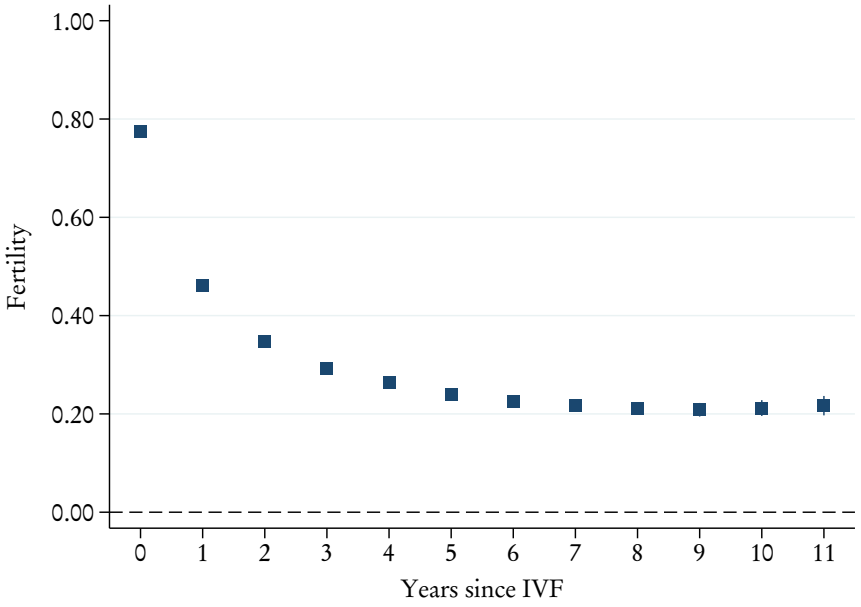
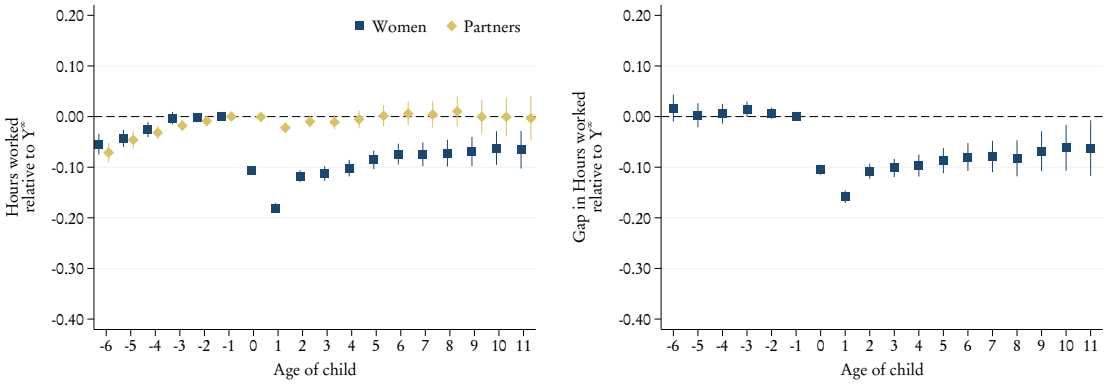


Figure A2. First stage. LPR-IV.

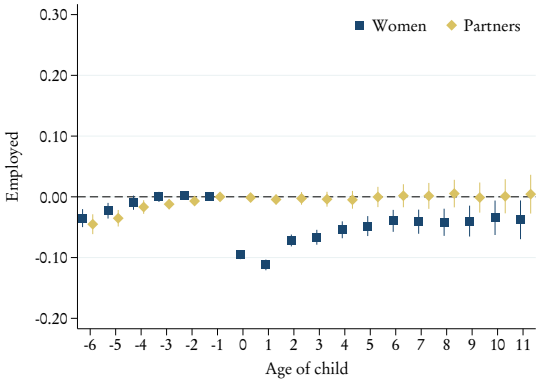
Note: First stage estimates using the IV model of Lundborg et al. (2017) as described in equation (4) on our data. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 4.4.

A.5 Other labor market outcomes

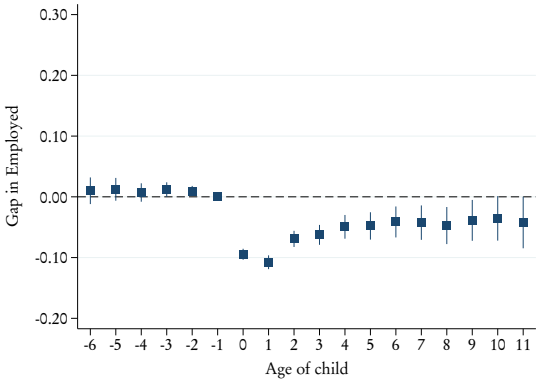


(a) Hours worked

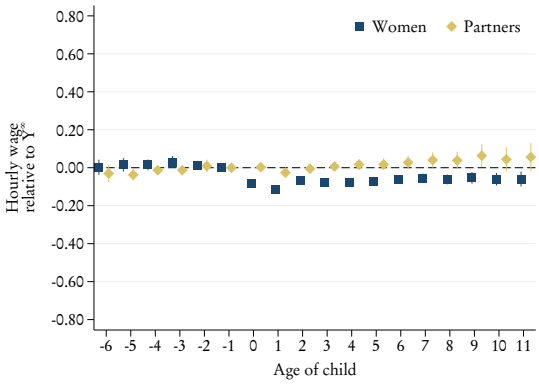
(b) Gap in hours worked



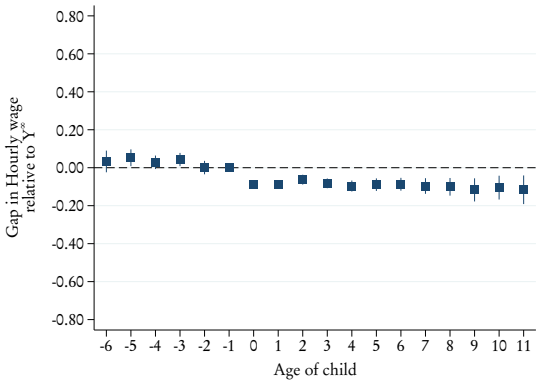
(c) Employment



(d) Gap in employment



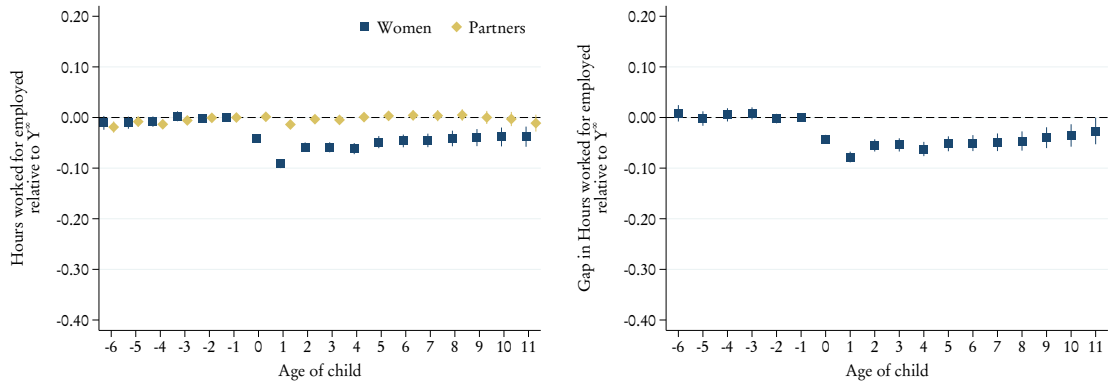
(e) Hourly earnings



(f) Gap in hourly earnings

Figure A3. Other labor market outcomes. Event.

Note: Event-study estimates from specification (3). The sample is all mothers (i.e., women who eventually have at least one child) undergoing IVF treatment and their partners. 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 4.4.

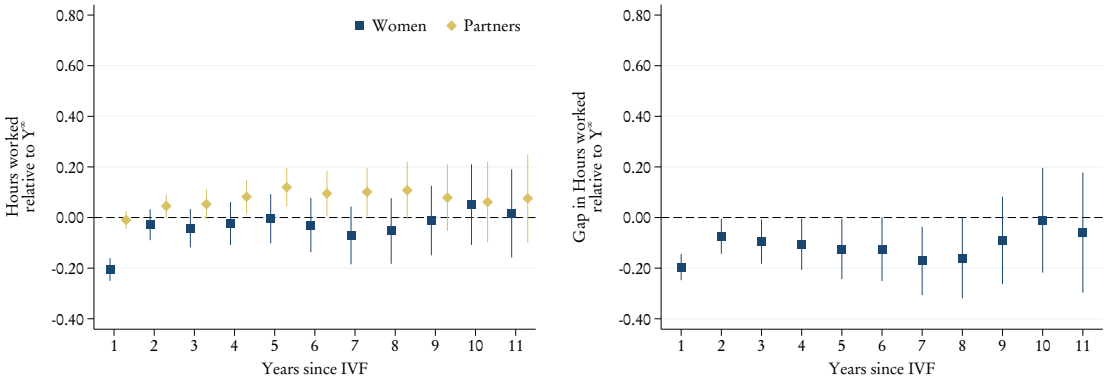


(a) Hours worked for employed

(b) Gap in hours worked for employed

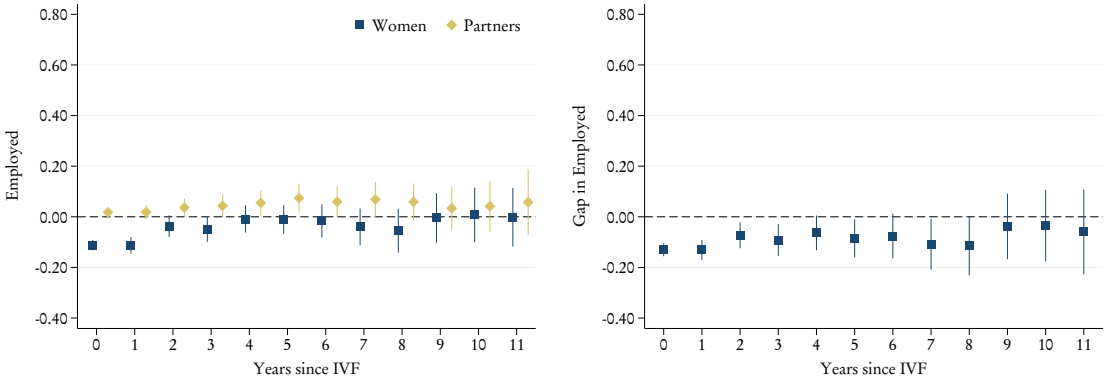
Figure A4. Hours conditional on employment. Event.

Note: Event-study estimates from specification (3). The sample is all mothers (i.e., women who eventually have at least one child) undergoing IVF treatment and their partners. 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 4.4.



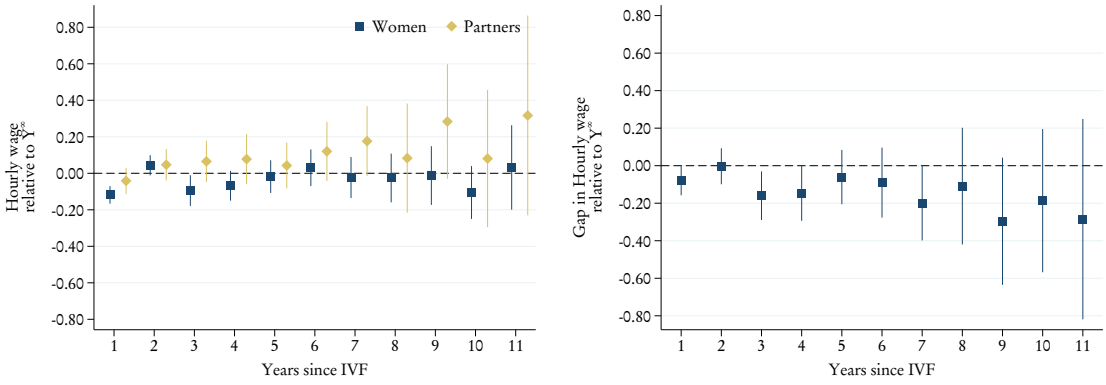
(a) Hours worked

(b) Gap in hours worked



(c) Employment

(d) Gap in employment

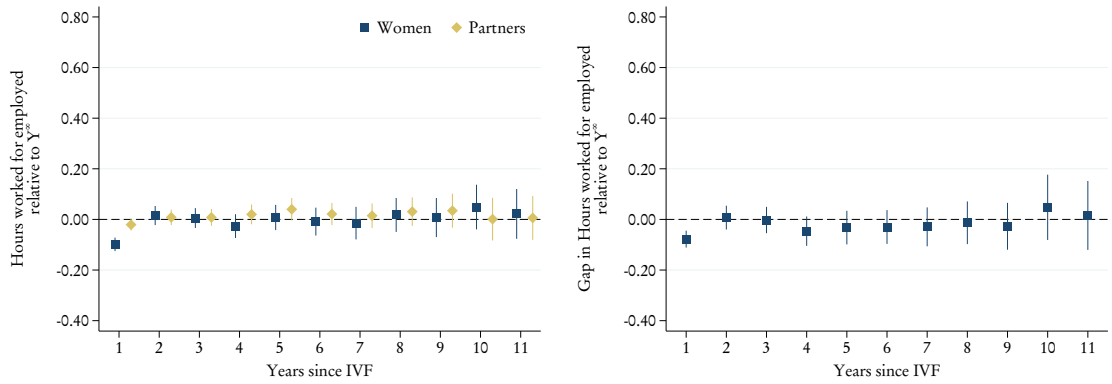


(e) Hourly earnings

(f) Gap in hourly earnings

Figure A5. 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. The sample is all women undergoing IVF treatment and their partners. 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 4.4.

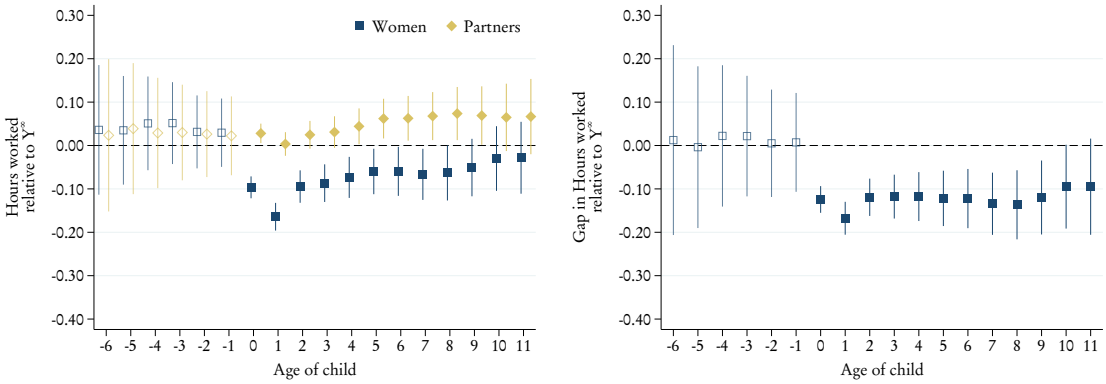


(a) Hours worked for employed

(b) Gap in hours worked for employed

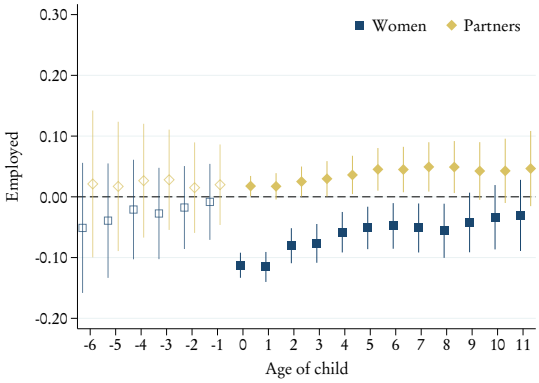
Figure A6. Hours conditional on employment. LPR-IV.

Note: Estimated effects of fertility using the IV model of Lundborg et al. (2017) as described in equation (4) on our data. The sample is all women undergoing IVF treatment and their partners. 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 4.4.

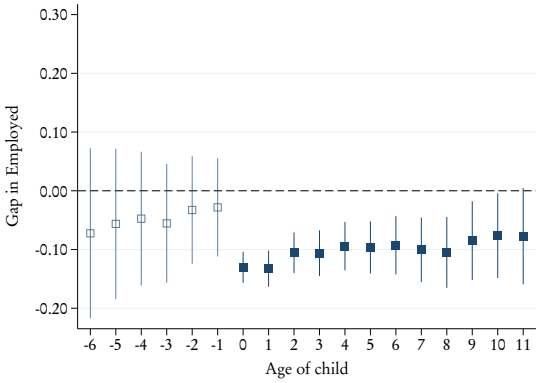


(a) Hours worked

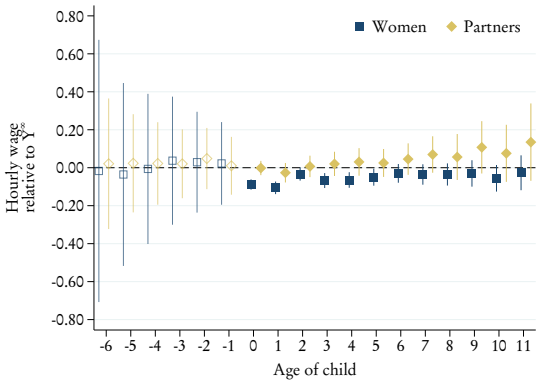
(b) Gap in hours worked



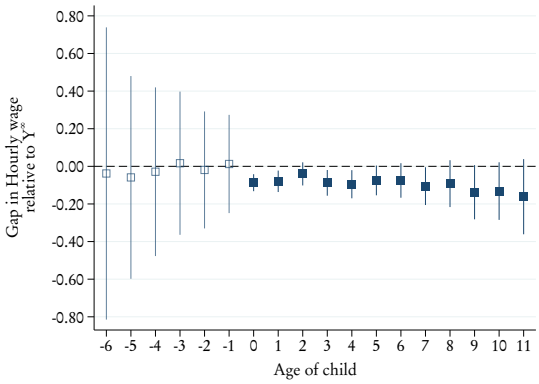
(c) Employment



(d) Gap in employment



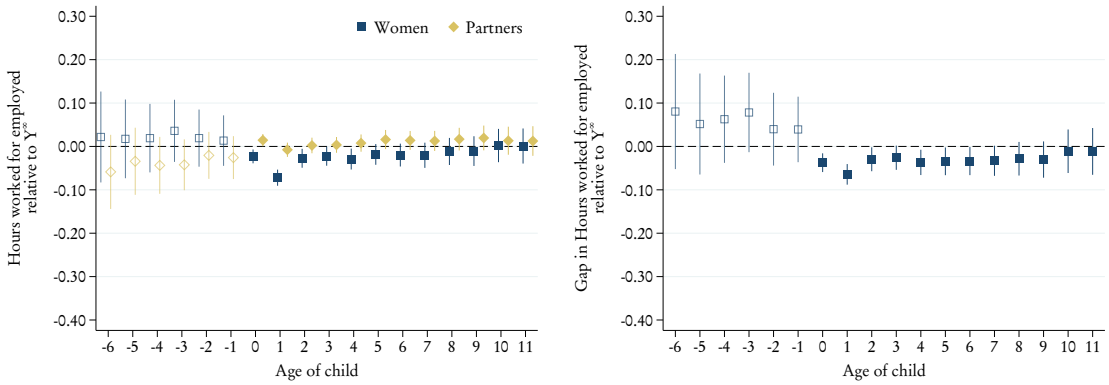
(e) Hourly earnings



(f) Gap in hourly earnings

Figure A7. Other labor market outcomes. Event-IV.

Note: Estimated effects of age of child using the event-IV model described in equation (7). The sample is all women undergoing IVF treatment and their partners. 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 4.4.



(a) Hours worked for employed

(b) Gap in hours worked for employed

Figure A8. Hours conditional on employment. Event-IV.

Note: Estimated effects of age of child using the event-IV model described in equation (7). The sample is all women undergoing IVF treatment and their partners. 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 4.4.

A.6 Non-labor market outcomes

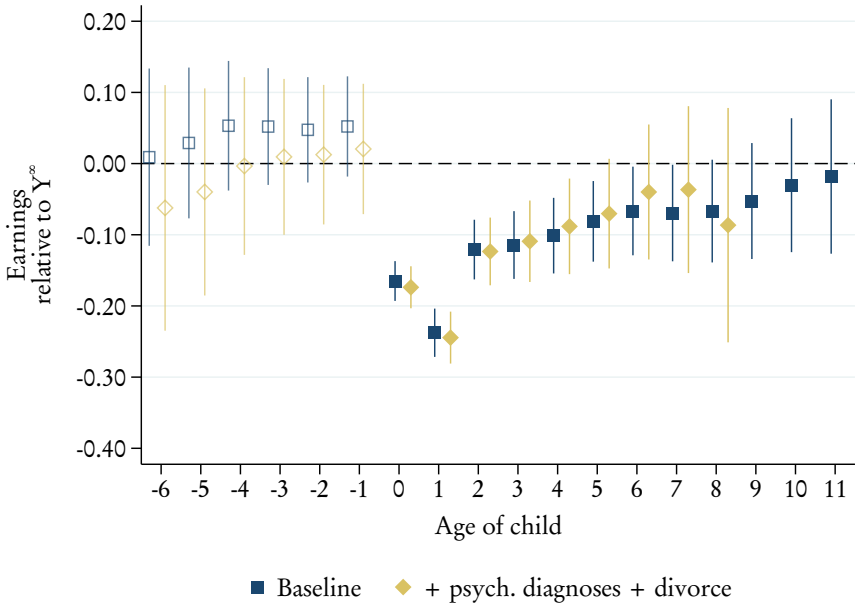
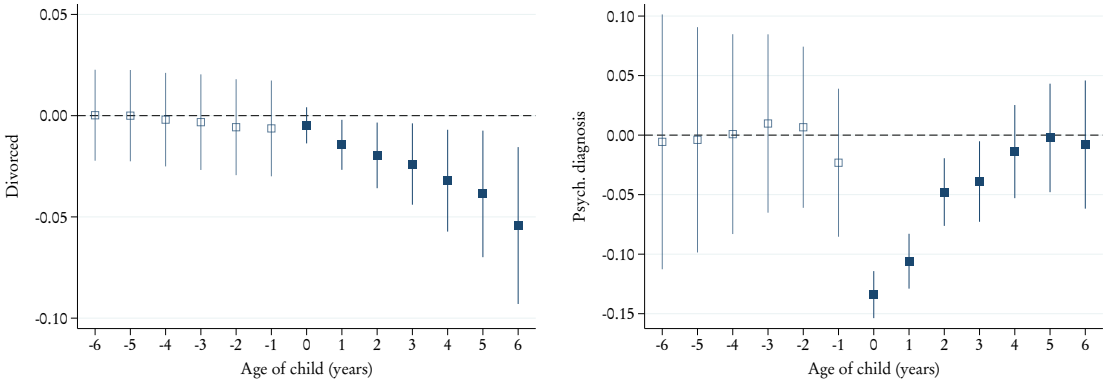


Figure A10. Robustness. Event-IV.

Note: Robustness checks of our event-IV model as specified in equation (7). Figure shows our baseline specification, alongside estimates that include controls for divorce and visits to a general practitioner for psychological symptoms. The sample is all women undergoing IVF treatment. All estimates are scaled relative to counterfactual earnings (Y^{∞}) as described in section 4.4.



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

Figure A9. Non-labor market outcomes – Event-IV.

Note: Results from our event-IV model shown in equation (7) using divorce and visits to GP for psychological symptoms as outcomes. The sample is all women undergoing IVF treatment.

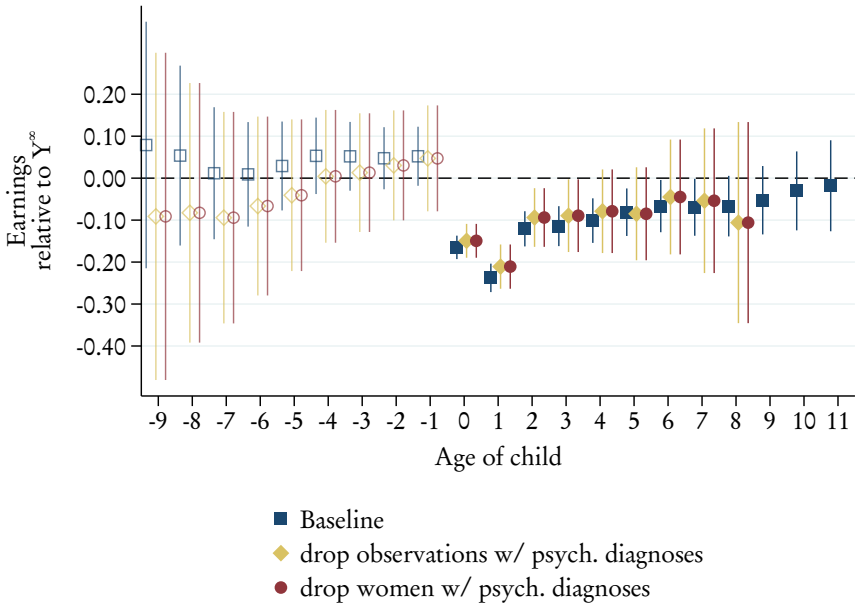


Figure A11. Robustness. Event-IV.

Note: Robustness checks of our event-IV model as specified in equation (7). Figure A11 show our baseline specification estimated in the sample of all IVF women, 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 4.4.

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
	(0.21)	(0.05)	(0.10)	(0.10)
Woman	0.07	-0.13	-0.02	-0.11
	(0.13)	(0.02)	(0.06)	(0.05)
Partner	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 (woman - partner), evaluated at $p = 11$. 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.

A.7 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 A12 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.

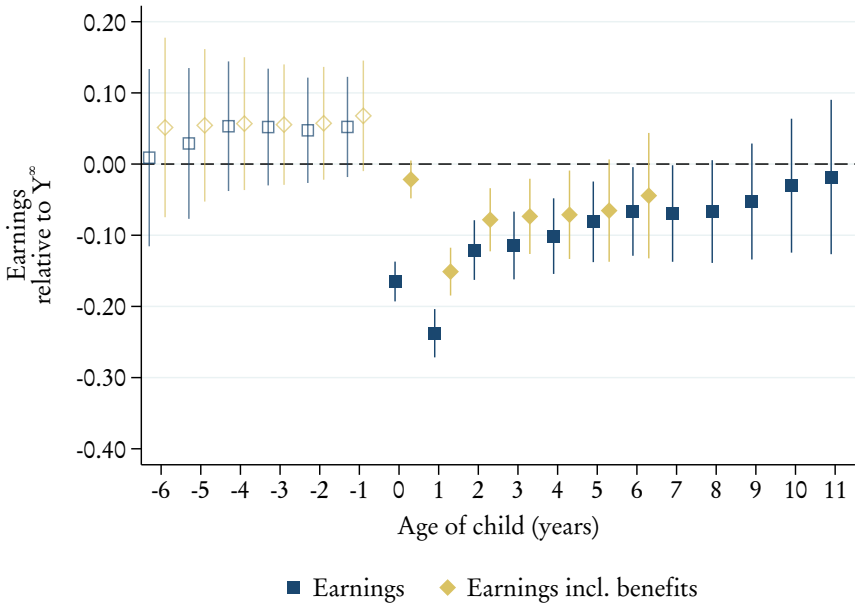


Figure A12. Earnings including benefits. Event-IV.

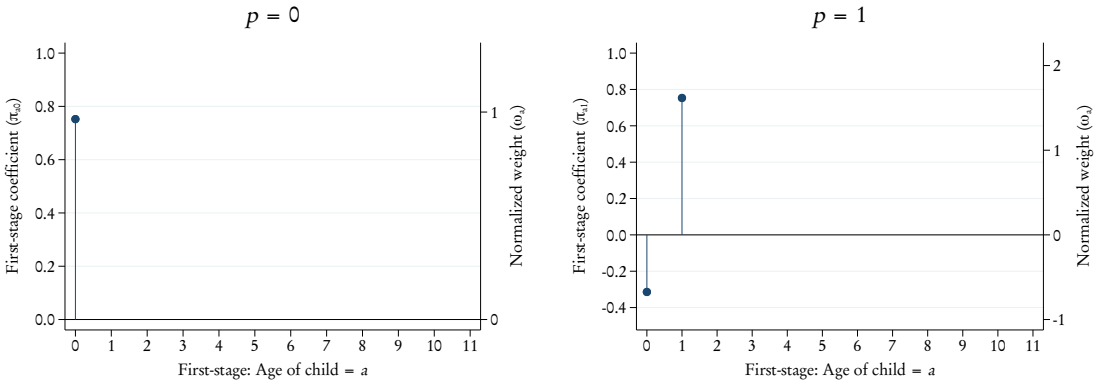
Note: Estimates from our 2SLS event study model as specified in equation (7). Outcomes are earnings and earnings including benefits. The sample is all women undergoing IVF treatment and their partners. Estimates are scaled relative to counterfactual earnings (Y^{∞}) as described in section 4.4.

A.8 Event-IV and LPR-IV

Table A8. First stage F-statistics for LPR-IV and event-IV.

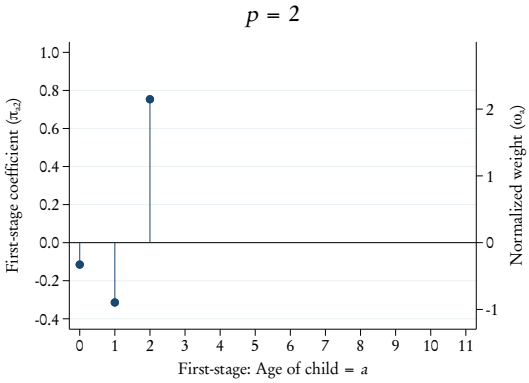
	(1)	(2)
Years since IVF (column 1) / Age of child (column 2)	LPR-IV F-statistic	Event-IV F-statistic
-6		669
-5		788
-4		838
-3		865
-2		932
-1		955
0	800	972
1	807	960
2	807	951
3	811	959
4	811	953
5	809	951
6	757	830
7	695	751
8	615	628
9	533	524
10	447	417
11	349	284

Note: F-statistics for first-stages from the LPR-IV model (equation 5) and the event-IV model (equation 7).

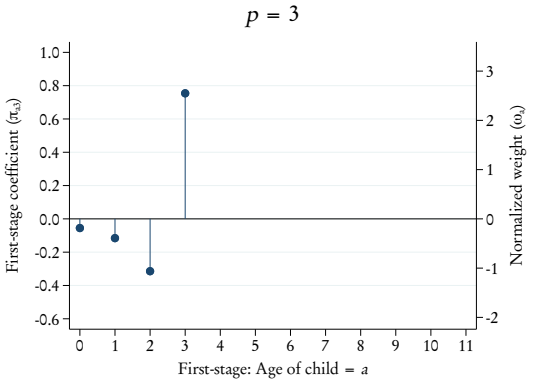


(a) Fertility weight at $p = 0$

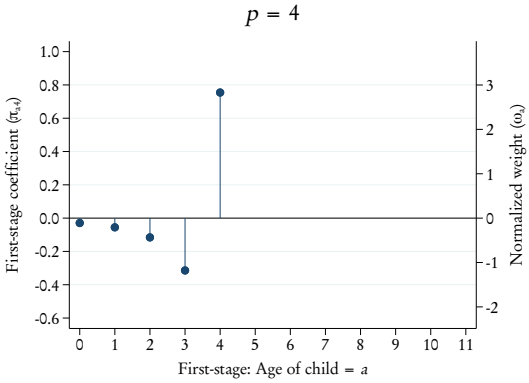
(b) Fertility weight at $p = 1$



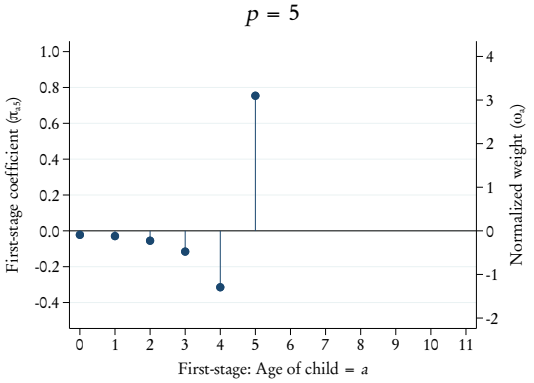
(c) Fertility weight at $p = 2$



(d) Fertility weight at $p = 3$



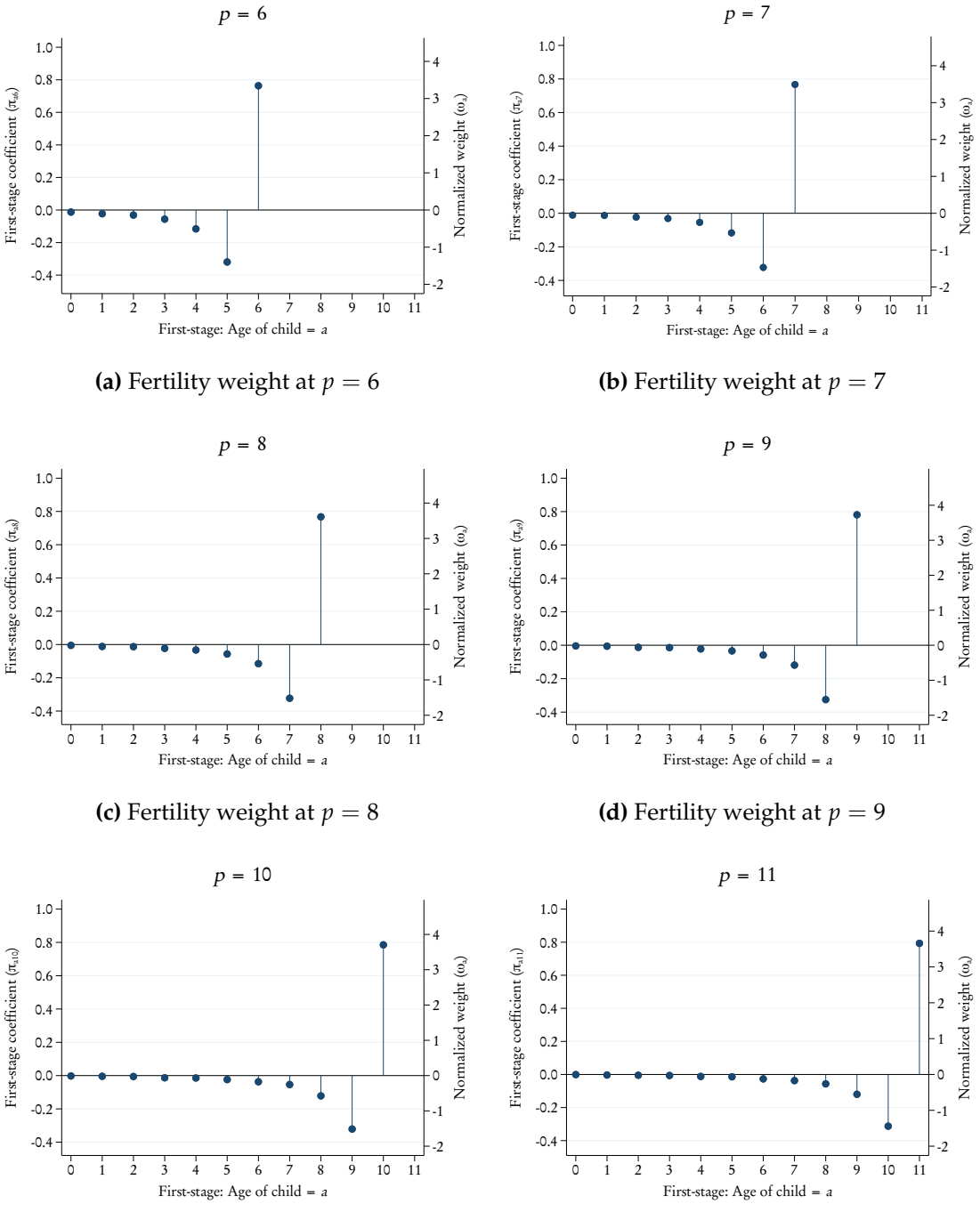
(e) Fertility weight at $p = 4$



(f) Fertility weight at $p = 5$

Figure A13. Fertility weights

Note: Fertility weights as defined in Section 7.1.



(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 A14. Fertility weights

Note: Fertility weights as defined in Section 7.1.

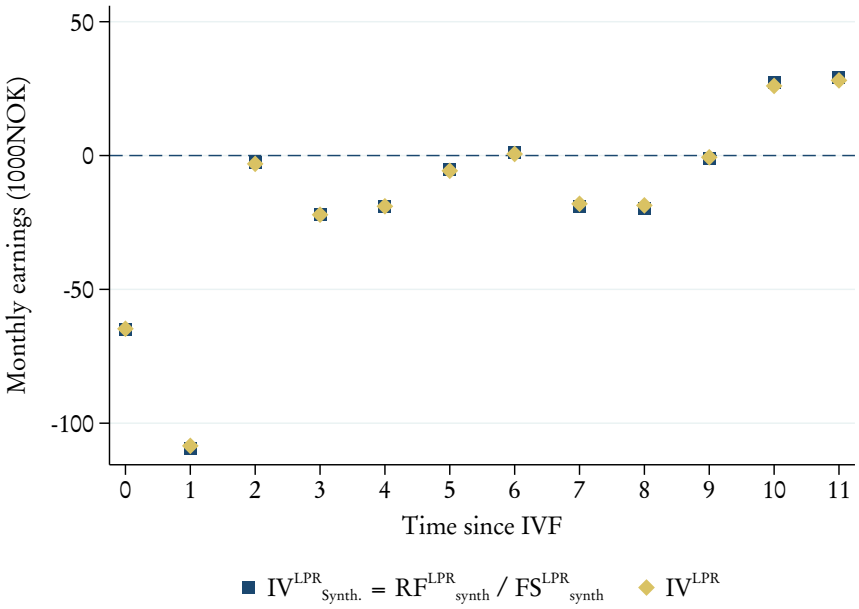


Figure A15. Combining results from LPR-IV and our event-IV model

Note: Figure shows results from our estimation of the IV model by Lundborg et al. (2017) alongside the rescaled event-IV estimates constructed from the reduced form and the rescaled first stages from our event-IV model in equation (7). The sample is all women undergoing IVF treatment. Estimates are scaled relative to counterfactual earnings (Y^{∞}) as described in section 4.4.

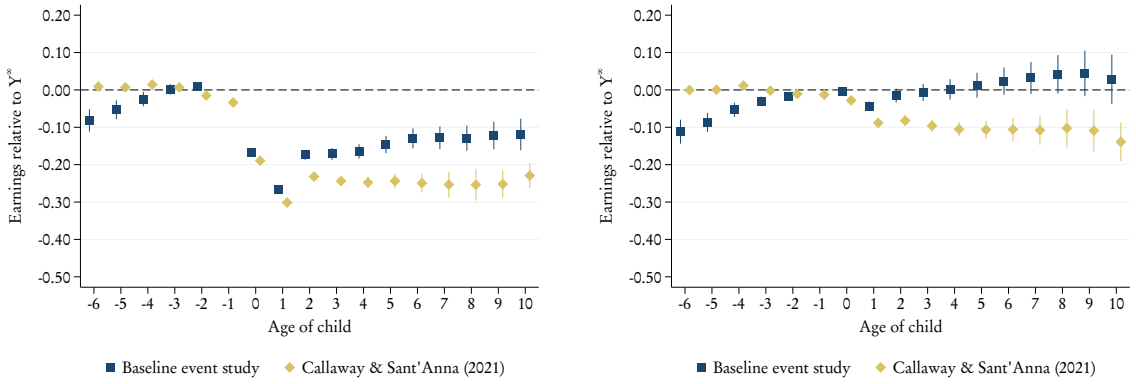
A.9 Complier characteristics

Table A9. Complier characteristics

	All IVF women		Compliers	
	Mean	Std.Dev.	Mean	Std.Dev.
Woman characteristics				
Age	33.00	(4.21)	33.12	(4.27)
Pre-IVF earnings	27.04	(16.95)	26.75	(17.07)
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	4.41	(10.21)	4.36	(10.14)
GP visits	0.37	(0.68)	0.38	(0.70)
Psychological symptoms	0.03	(0.18)	0.04	(0.20)
Hospital days	0.90	(5.17)	0.78	(5.73)
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	36.73	(24.50)	36.57	(24.28)

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.

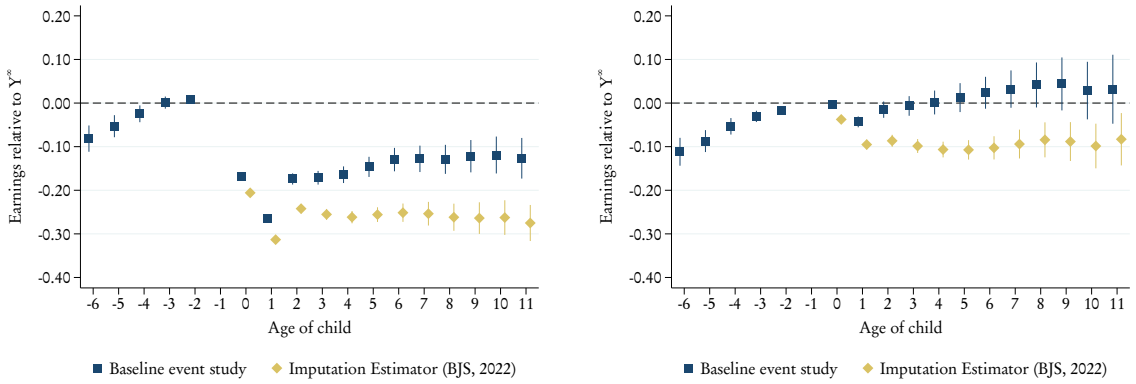
A.10 Alternative event-study estimators



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

Figure A16. Results based on the treatment-cohort Callaway and Sant'Anna (2021) estimator

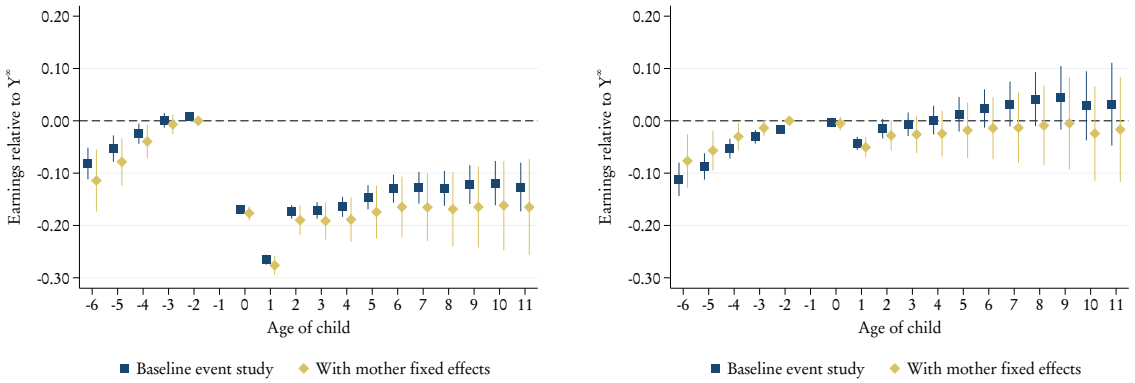
Note: This figure shows the estimated results from the event model using the conventional estimator as applied in f.e. Kleven et al. (2019) and results using the estimator proposed by Callaway and Sant'Anna (2021) with bootstrapped standard errors, allowing for effect heterogeneity by mother's age at birth. The sample is all mothers (i.e., women who eventually have at least one child) undergoing IVF treatment and their partners.



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

Figure A17. Results based on the Borusyak et al. (2024) imputation 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 Borusyak et al. (2024) with bootstrapped standard errors, allowing for effect heterogeneity by mother's age at birth. The samples consist of all mothers (i.e., women who eventually have at least one child) undergoing IVF treatment and their partners.



(a) Woman's earnings

(b) Partner's earnings

Figure A18. Results based on event-study specification with mother fixed-effects.

Note: OLS event study estimates from specification (3), 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 4.4. The samples consist of all mothers (i.e., women who eventually have at least one child) undergoing IVF treatment and their partners.