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

Simon Bensnes† Ingrid Huitfeldt‡ Edwin Leuven§

May 26, 2023

Abstract

This paper presents novel methodological and empirical contributions to the child penalty literature. We propose a new estimator that combines elements from standard event study and instrumental variable estimators and demonstrate their relatedness. Our analysis shows that all three approaches yield substantial estimates of the long-term impact of children on the earnings gap between mothers and their partners, commonly known as the child penalty, ranging from 11 to 18 percent. However, the models not only estimate different magnitudes of the child penalty, they also lead to very different conclusions as to whether it is mothers or partners who drive this penalty – the key policy concern. While the event study attributes the entire impact to mothers, our results suggest that maternal responses account for only around one fourth of the penalty. Our paper also has broader implications for event-study designs. In particular, we assess the validity of the event-study assumptions using external information and characterize biases arising from selection in treatment timing. We find that women time fertility as their earnings profile flattens. The implication of this is that the event-study overestimates women's earnings penalty as it relies on estimates of counterfactual wage profiles that are too high. These new insights in the nature of selection into fertility show that common intuitions regarding parallel trend assumptions may be misleading, and that pre-trends may be uninformative about the sign of the selection bias in the treatment period.

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.

†Statistics Norway.
‡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 affecting labor market outcomes. Neglecting these confounding factors results in omitted variable bias. To address this issue, the recent literature has primarily 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 et al. (2017), and later used by Gallen et al. (2022), is to use IVF (in vitro fertilization) as an instrumental variable for fertility. They demonstrated that, given participation, the outcome of IVF treatment is as good as random, and hence, 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 unspecified age) at a given point in time since the IVF attempt. Their 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 a consequence, the resulting estimated treatment effects are latent mixtures of different dynamic treatment effects and thus not directly comparable to event-study estimates. Moreover, as also pointed out by Lundborg et al. (2017), if the impact of having children on female labor earnings is particularly large when children are young, then the fertility response is underestimated (a positive bias) because the always-taking mothers in the comparison group have younger children.[1]

In this paper we re-examine the labor market effects of having children using admin-

[1] This is technically a violation of the exclusion restriction because IVF success not only affects fertility but also the age of the child.

istrative 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 a novel event-study model that combines these approaches (referred to as event-IV). The advantage of the event 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 success in IVF treatment in an instrumental variable setup.

In our empirical analysis 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, but varying increase in the long-term earnings gap between parents, commonly known as the *child penalty*. The event-study model estimates the earnings gap to be 18 percent, whereas the LPR-IV model estimates it at 11 percent. The event-IV model falls in between at around 13 percent. While the spread in the estimated child penalty warrants attention in itself, the key policy implications of these models rest on whether it is the mother or the partner who drives results. If the child penalty 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 the penalty results from a reduction in women's earnings, such policies may help achieve equal pay.

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 only about one fourth. More specifically, the event-study model indicates large negative long-run effects on maternal earnings of around 16 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 again in between, estimating a reduction of only 3 percent. Turning to partners' earnings, the ordering of the estimates goes in the opposite direction: The event-study model estimates a nonsignificant decrease of 1 percent, while both the LPR-IV model and the event-IV model suggests a substantial increase of around 10 percent.

We explore the sources of bias and differences in estimates between models. First, we demonstrate that the estimates from the event-IV model map into those from the

LPR-IV model as fertility is the same as having a child of any age. Using the derived weights from this mapping, we show how the LPR-IV model provides estimates that are mixtures of the effects of having children of various ages, and that with time the model puts increasingly negative weight on the effect of children born after the first IVF trial.

Next, we find that already half of the difference between the standard event-study model and the event-IV estimates of the effect on mothers' long-run earnings is explained by adjusting earnings profiles for time since the IVF trial (a predetermined variable). The remaining difference between the event-study estimates with these timing controls and our event-IV estimates is explained by fertility from natural conception and adoption (and addressed by the instrumental variable).

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 we find that adjusting for a linear extrapolation of the pre-trend exacerbates 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., 2022; Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020). We implement the imputation estimator of Borusyak et al. (2022) and obtain even more pronounced negative effects on maternal earnings than in the standard event study specification, which is consistent with the results based on the extrapolation of pre-trends.[2]

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. Hotz et al., 2005; Cristia, 2008; Aguero

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

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 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 IVF 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. Section 8 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

[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.

to go through several rounds of inserting fertilized eggs within one treatment. In our analyses we refer to trials or attempts as the *insertion* of eggs, which is identified in the data since hospitals are reimbursed by the government for each such procedure. 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 three treatments at a public hospital is about NOK 18 000 (USD 2 000 in 2019) 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

Any effect of fertility on earnings and labor supply is channeled through the labor market and the social insurance and benefit system. Since the 1970's the Norwegian government has gradually introduced several major support systems for parents (NOU 2017:6, 2017). In the time period we study both parents had the legal right to a total of almost one year of parental leave following the birth of a child. Parents could choose between slightly less than one year of parental leave with 100% wage replacement, and a ten week longer period with 80% wage replacement.[5] Additionally, women could apply for welfare support during pregnancy if their working conditions could be harmful to the health of fetus or the mother. Employers were (and are) legally bound to not discriminate based on pregnancy when hiring, promoting, or firing employees. Further, sick leave benefits were quite generous such that workers have the legal right to a certain number of sick leave days both when they are sick themselves and to care for sick children. Last, the national government expanded the formal child care sector substantially starting in the early 2000's such that nearly all children could attend subsidized child care if the parents wished so (Andresen and Havnes, 2019; Drange and Havnes, 2019). In sum there were different support systems at various stages of pregnancy, birth, and child upbringing that could compensate for earnings lost due to fertility.

[5]Part of the parental leave is reserved for the mother, and part for the father. Eligibility is contingent on sufficient income in the year prior to birth. The eligibility criteria are quite lenient and most parents qualify.

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 contain 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 2017.

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 onwards, 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 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 diagnosis code. The data are available for us from 2006 to 2017.

Our main labor market outcomes are derived from the employer-employee registry. This registry contains information on start and stop dates of a job spell, as well as the corresponding labor income, occupation, sector and contracted hours.[6] We have access to these data for the period 2004 to 2017.[7]

[6] 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.

[7] 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.

We define four variables to capture individuals' labor market attachment. Our main outcome, *Earnings*, captures the yearly labor income. *Employed* is a binary indicator equal to one if the individual has positive labor income in a given year, zero otherwise. *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 leave estimates for the other outcomes to the appendix.

Sample

Our main 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 have excluded 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.[8] 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 IVF women (2009 to 2017), and who were registered with a partner in the year of conception.

Descriptive statistics are presented in table 1. Column (1) focuses on the sample of IVF women, while column (2) describes 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).[9] 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.[10] In our sample, the average number of IVF trials is about 2.8, and the end-of-period success rate is 61 percent. In total 76 percent of the IVF women eventually have at least one child. The difference between realized

[8] 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.

[9] 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.

[10] 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.

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

	IVF (1)	Non-IVF (2)
Mother characteristics		
Number of IVF attempts	2.84	
Any success	0.61	
Fertility, endpoint	0.76	1.00
Total number of children	1.14	1.59
1 children	0.42	0.49
2 children	0.30	0.44
3 children	0.04	0.07
4 children	0.00	0.00
Age	31.8	28.4
Education		
- Compulsory	0.14	0.17
- High School	0.24	0.23
- Bachelor	0.42	0.41
- Master	0.20	0.19
Earnings (1000 NOK)	362.7	289.9
Hours (FTE)	0.88	0.79
Employed	0.88	0.85
Hourly earnings	221.2	198.7
Sickness absence days	15.0	11.1
Visits to general practitioner (GP)	2.51	2.16
Visits to GP registered with psychological diagnosis	0.14	0.12
Hospital days	2.13	1.01
Partner characteristics		
Age	35.1	31.2
Female	0.01	0.01
Education		
- Compulsory	0.17	0.20
- High School	0.39	0.37
- Bachelor	0.27	0.26
- Master	0.17	0.17
Earnings (1000 NOK)	454.9	385.4
Hours (FTE)	0.84	0.78
Employed	0.87	0.84
Hourly earnings (NOK)	281.8	258.7
N Women	10 033	109 791

Notes: Column (1) shows descriptive statistics for women who had at their first child without IVF treatment during the period 2009 to 2017. Column (2) shows descriptive statistics for women who had at least one IVF trial who had at least one IVF trial over the period 2009 to 2016. 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[8] 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.

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, 42 percent of the IVF women have one child, and 30 percent have two children, 4 percent have three children, and virtually none have four children or more. Fifty-five percent (0.42 / 0.76) of IVF mothers have one child. For comparison, non-IVF mothers are more likely to have two or more children; 49 percent have one child, while 44 percent have two children, and 7 percent have three or more children.[11]

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 in the two samples, with 42 percent of the IVF women holding a bachelor's degree and 20 percent holding a master's degree, compared to 41 and 19 percent in the non-IVF sample. IVF women have higher earnings and work more hours compared to non-IVF mothers. 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 IVF women, 88 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 199 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 spent 2.1 days per year at the 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 that were coded with a psychological diagnosis; the average number of such visits was 0.14 per ÿear 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.[12]

Compared to non-IVF mothers giving birth during the same period, we therefore

[11] 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).

[12] Statistics Norway defines a full-time position as equaling 1,950 hours per year.

see that IVF women tend to be somewhat older, and earn and work more, while their educational attainments are similar. 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 somewhat more likely to have more than one child, their final fertility patters are overall very 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 potential outcomes

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

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. By τ_t we denote flexible controls for time through calendar year dummies. The coefficients δ_a allow for age-of-child specific shifts in the outcome.

Observed outcomes map into potential outcomes as follows

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

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 $\epsilon_{it} \equiv \epsilon_{it}^{\infty} + \sum_{a \geq 0} \mathbb{1}_{\{\text{age } \mathbf{1st} \text{ child}_{it}=a\}} \epsilon_{it}^{a}$. Equation (1) corresponds to a standard event-study specification.

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:

$$
y_{it} = \sum_{a \neq -1} \delta_a \mathbb{1}_{\{\text{age child}_{it}=a\}} + x_{it}'\phi + \tau_t + \epsilon_{it} \qquad (2)
$$

where for notational convenience negative values of a refer to time before birth.[13] 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 to seven years after birth.

Assuming no heterogeneity in treatment effects, the counterfactual outcome profile in (2) is identified from the pre-birth wage profiles, and identification of the child penalties δ_a 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 child penalty.

The existing child penalty literature often focuses on the earnings difference between mothers and fathers, which allows for a weaker exogeneity assumption than required by (2). As we make precise below in section 4.4, rather than assuming that women who have children at different times would have the same earnings development in absence of children, the assumption becomes that the earnings difference between mothers and fathers would develop similarly in absence of children.

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 which 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

$$y_{ip} = \gamma_p \text{Fertility}_{ip} + x'_{ip}\psi_p + \theta_p + u_{ip} \tag{3}$$

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

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

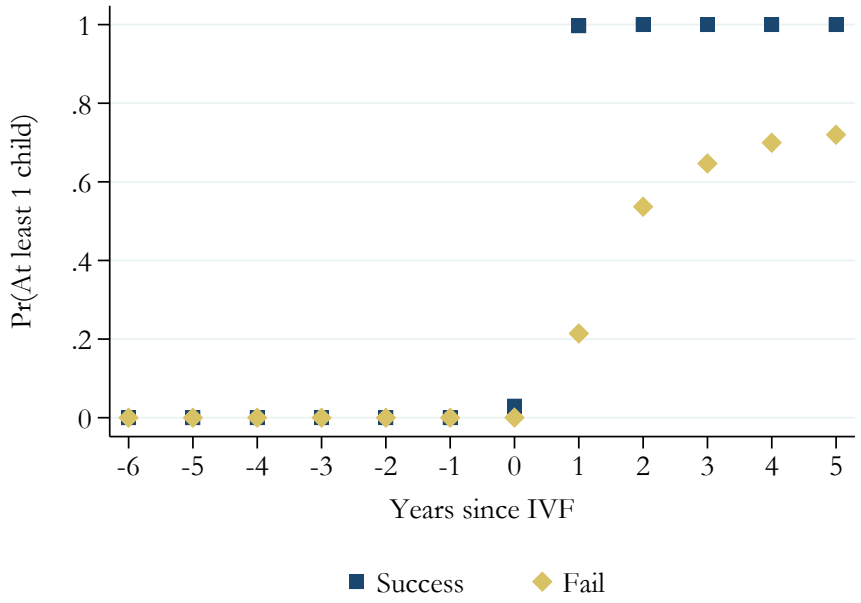


Figure 1. Fertility by success at first IVF trial

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

is zero otherwise. We follow Lundborg et al. (2017) and include controls for mother's age, x_{ip}, and dummies for calendar year. Because success correlates with mothers' education we interact x_{ip} with the education level at the IVF trial to make sure that the instrument is conditionally exogenous.[14] In addition, equation (3) 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:

$$\text{Fertility}_{ip} = \pi_p \text{success}_i + x'_{ip}\lambda_p + \mu_p + w_{ip} \tag{4}$$

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

[14] Lundborg et al. (2017) also control for average earnings in the years leading up to the first trial. We discuss using pre-treatment earnings as a control in appendix section A.6.

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 70 percent of the women with a failed first IVF realizes motherhood.

In practice many women therefore end up having children despite a failed 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 (4) 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 child penalty is larger when children are young then the fertility estimates γ_p will be a mixture of child penalties 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.

4.3 Event-IV

The advantage of the standard event-study setup (2) is that it recovers well defined child penalties, 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 mixtures of child penalties. 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 and not a latent mixture of treatments arising from differential compliance over time (delayed fertility) and therefore addresses this potential violation of exclusion.

Second, 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 (2):

$$y_{it} = \sum_{a \geq 0} \delta_a \mathbb{1}_{\{\text{age child}_{it}=a\}} + x'_{it}\phi + \tau_t + \sum_p \gamma_p \mathbb{1}_{\{\text{time since IVF}_i=p\}} + \epsilon_{it} \tag{5}$$

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 (5) 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

$$\begin{aligned}\mathbb{1}_{\{\text{age child}_{it}=a\}} = &\sum_p \pi_{ap} \mathbb{1}_{\{\text{time since IVF}_i=p\}} \times \text{success}_i \\ &+ \sum_p \theta_{ap} \mathbb{1}_{\{\text{time since IVF}_i=p\}} + x'_{it}\tilde{\phi}_a + \tilde{\tau}_{at} + u_{iat} \end{aligned} \tag{6}$$

While the event-study specification of equation (2) 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.

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.

4.4 Definitions of the child penalty

For ease of interpretation and comparability across contexts we focus on relative rather than absolute effects in cardinal units such as Norwegian Kroner. 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/fertility. For women the estimand is therefore the following

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

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 child penalty δ_a from the observed average wage of these women:

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

For the IV approach we estimate the counterfactual outcome following Abadie (2003). The implementation with linear 2SLS involves re-estimating the child penalty for each a where the outcome variable equals $-\mathbb{1}_{\{\text{age child}_{it} \neq a\}} y_{it}$.[15] If we denote the resulting counterfactual outcome for mothers by $\delta_a^{\infty,women}$ then the rescaled IV child penalty 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

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

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 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:

$$P_a = \frac{\delta_a^{women}}{\delta_a^{\infty,women}} - \frac{\delta_a^{men}}{\delta_a^{\infty,men}} \tag{7}$$

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 child penalties of mothers and

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

fathers:[16]

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

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 joint 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.[17] As in the 2SLS specification in equation (5) and (6), all regressions include controls for calendar time, time since IVF treatment, and maternal age, which is a known predictor of success (CDC, 2012). 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 higher educational attainments have higher earnings; and women with poorer health, as measured by visits to their primary care physicians, and visits to primary care physicians resulting in a psychological diagnosis, 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.

Exogeneity of IVF success requires the variables that are correlated with pre-treatment earnings in the first column to be uncorrelated with our instrument. Column (2) indicates that these characteristics are generally not predictive of the instrument. For example, while hospital days is marginally associated with the IVF success rate, it is not

[16] 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.

[17] 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)	
Mother 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 diagnosis	-0.126	(0.034)	0.006	(0.010)
Hospital days (/10)	0.000	(0.002)	-0.001	(0.001)
Education (ref. master)				
- Compulsory	-1.341	(0.058)	-0.064	(0.019)
- High School	-0.819	(0.053)	-0.022	(0.016)
- Bachelor	-0.491	(0.049)	0.003	(0.014)
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		0.32	
Joint F [p-value]	117.5 [< .001]		3.3 [< .001]	
Joint F [p-value] excl. mother education	44.3 [< .001]		1.3 [0.228]	
N Women	10 033		10 033	

Note: This table reports estimates 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 (5) and (6), both regressions include dummies for calendar time, time relative to IVF treatment, and mother age. Joint Fs [p-value] refer to tests of joint significance of the characteristics shown in the table.

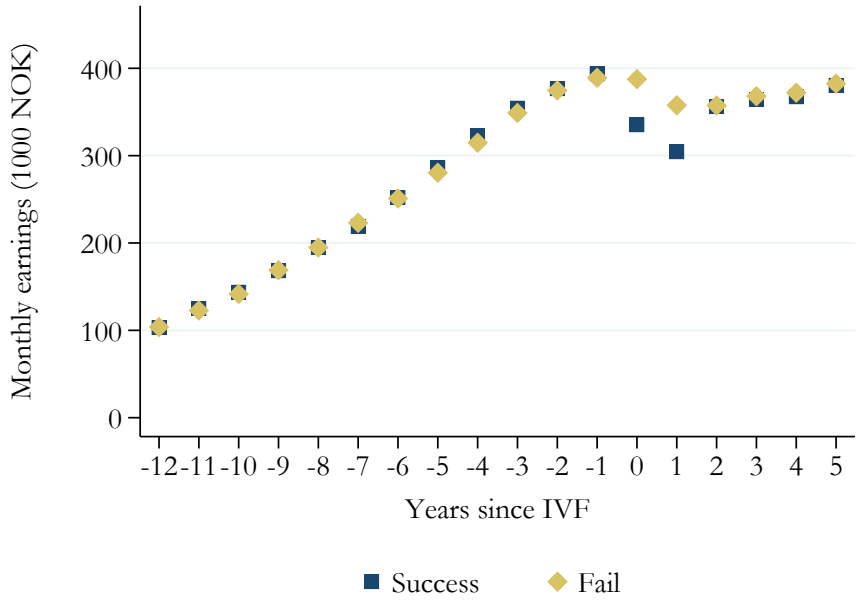


Figure 2. Conditional independence of IVF success

Note: This figure plots average earnings for each year relative to the IVF trial. Average earnings are computed separately by IVF success, calendar year, maternal education, maternal age and then plotted as the average for each year since the IVF trial.

predictive of earnings, and hence not a potential confounder. The regression also reveals that maternal education is predictive of success: the success rate is 6.4 percentage points lower for women with only compulsory education relative to women with a master's degree. This is in line with Groes et al. (2017) who, using Danish data, also find that success correlates with education. As a result, all variables are jointly significant in explaining success. However, a test for joint significance of all variables except mother's educational attainment is not significant and renders a p-value of 0.23. To avoid any potential bias arising from differences across women with different educational background, we allow all control variables in the IV specifications to vary by education.

While 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, conditioning on time-since-IVF, calendar time, maternal age, and maternal education. More precisely, we compute average earnings separately by IVF success, calendar year, maternal education, maternal age and then plot the average 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. Nonetheless, in the robustness analyses below we adjust for pre-IVF earnings and find that the imbalance does not introduce

any meaningful bias.

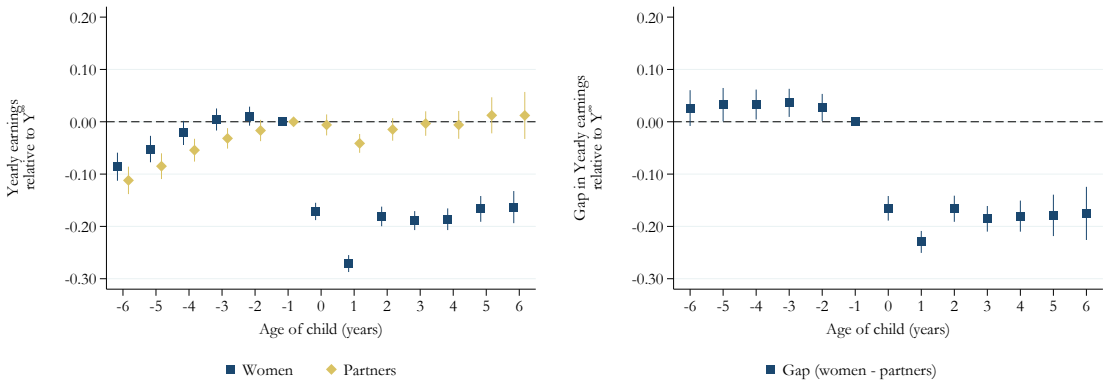
6 Children and labor market outcomes

We now present the estimated effects on earnings for the three different models described in Section 4: the standard event-study, the instrumental variable effect estimates of fertility since the IVF attempt (LPR-IV), and our specification that combines these two approaches (event-IV). For each model, we report estimates for mothers, partners, and the difference between the two (as in equation 7).

6.1 Event-study estimates

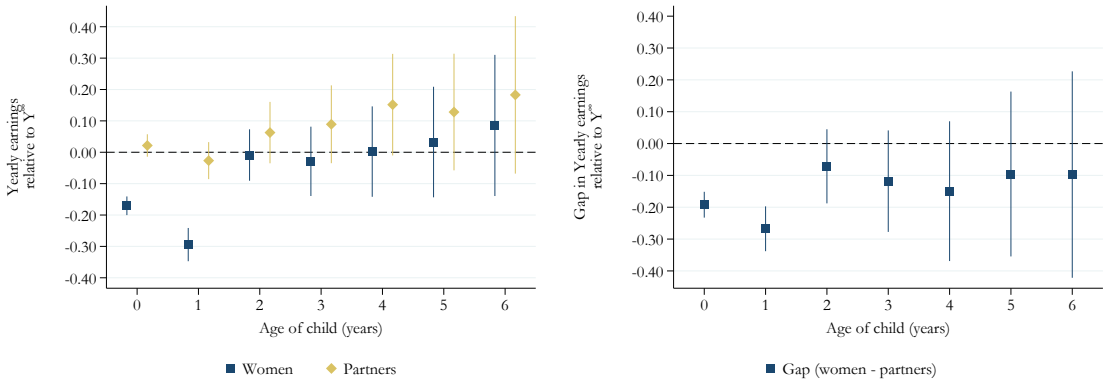
We start by reporting the results using the regular event-study specification of equation (2), estimated on IVF-mothers and their partners in figure 3(a).[18] Both women and partners display a comparable pre-trend leading up to birth, indicating that women who have children earlier are on relatively 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 16 percent in the longer run. Partners, in contrast, experience almost no change in earnings following childbirth.

[18] 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) Earnings **(b)** Gap in earnings

Figure 3. Earnings. Event study.
Note: OLS event study estimates from specification (2). Panel (a) shows effects separately for women and partners, panel (b) shows the difference between women and partners. Estimates are scaled relative to each gender's counterfactual earnings (Y^{∞}), as described in section 4.4. Samples are mothers and partners undergoing IVF treatment.



(a) Earnings **(b)** Gap in earnings

Figure 4. Earnings. LPR-IV.

Note: Estimated effects of fertility on earnings using the LPR-IV model described in equation (3) 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.

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

These results confirm other findings from Norway and estimates for comparable countries like Denmark which find similar pre-trends and effects for mother, partners and differences on the longer run. In appendix figure A1 we confirm this by showing that we obtain very similar results when we estimate the same event specification on the sample of non-IVF women. Differences are even smaller when we reweight the IVF-sample by the age and education of non-IVF women as discussed in more detail in appendix section A.3 and shown in appendix figure A2.

6.2 LPR-IV estimates

We now present the estimated earnings effects of fertility using the LPR-IV model described in equation 3 with the outcome of the IVF treatment as the instrumental variable. Appendix figure A3 reports the estimated first stages from equation (3), essentially the difference between the average fertility rates between successful and failed IVF attempts shown in figure 1. By construction, the first stage equals one 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 0.3 in the longer run. Despite this decline, the estimates are all highly significant: Women who are successful in their first IVF-trial are therefore always more likely to have chil-

dren than those who failed their first trial. The F-statistic is well above conventional levels in each year since IVF and are reported in appendix table A2.

Figure 4(a) shows the IV estimates of equation (4), separately for women and their partners. Women's earnings drop to about 30 percent in the year following the IVF treatment, but the effect quickly reverts to zero in the third 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.

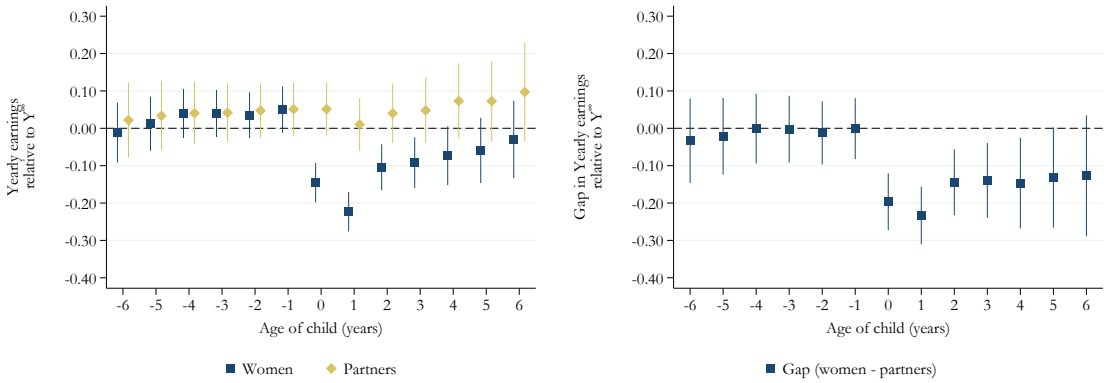
In contrast, partners see no earnings drop immediately following IVF treatment. If anything, there is a small earnings premium in the longer run. Between the second and sixth year after undergoing IVF, during which time earnings seem to remain relatively stable, partners experience an average increase in earnings of around 11 percent. Although the yearly estimates are imprecise, the average over these five years is significant at conventional levels.

Figure 4(b) reports the estimated effect of fertility on the earnings gap between women and their partners. This fluctuates a bit over time, averaging at 11 percent in the longer run (again over years 2 through 6 after the IVF trial), 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 women's long-run penalties in the neighborhood of 18 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 6. F-statistics for the first stages are reported in appendix table A2 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 22 percent, the long run earnings penalty is around 3 percent. This is a small fraction of the penalty estimated in the event-study model and the differences are statistically different at conventional significance levels. No earnings drop around childbirth is seen for partners. In contrary,



(a) Earnings **(b)** Gap in earnings

Figure 5. Earnings. Event-IV.

Note: Estimated effects of age of child on earnings using the event-IV model described in equation (6). 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 estimates suggest an increase in earnings over time, reaching around 10 percent in the long run. Figure 5(b) plots the estimated gap between women and partners from the event-IV model. There is no evidence of an earnings gap before birth, at which point it drops to around 23 percent, before stabilizing at around 13 percent in the longer run. This long run parental earnings gap is primarily driven by the partners.

For completeness, appendix figure A6 reports results for additional labor market outcomes (hours, employment and hourly wages) for all three models. The broad takeaway from the event-IV model is that the results look very similar across all labor market outcomes, but we lack precision to separate the different channels. We find that the gap in both employment and hours worked increase after parenthood. We refer to appendix A.5 for a full discussion of these results.

Robustness of event-IV model In appendix A.2 we explore several potential challenges to the internal validity of our results. First, the outcome of an IVF trial may impact women directly through other mechanisms than fertility. A potential worry could be that failure to conceive may impact mental health or divorce risk which in turn affect labor market outcomes. Such mediation would violate the exclusion restriction. In appendix figure A7 we discuss and analyze the role of such potential mediators finding that there is little reason to believe they bias our findings. Second, estimated effects for women and partners both display a minor imbalance before birth. In appendix figure A8 we show that controlling for income earned in the years before the IVF trial as in Lundborg et al. (2017) removes all evidence of any bias. Finally, there are several ma-

jor welfare programs in Norway that aim to replace lost labor market earnings such as parental and sick leave. Our preferred 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 A9 shows that this, as expected, dampens the estimates in the very short run, but does not impact 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.[19] 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, and column (4) shows the difference between the estimates from the event model and the event-IV model. This difference can be interpreted as the bias present in the event estimates under the assumptions of the event-IV model and absent notable complier heterogeneity which we document below.

The first thing to note is that the estimates of the fertility impacts on the earnings gap between mothers and partners are sizable in the three different models. First, consider the estimates from the LPR-IV and the standard event-study model. The LPR-IV model estimates a long-run impact on the parental earnings gap of almost 11 percent, compared

[19] The event model and the event-IV model are evaluated at child age $a = 6$. Point estimates provided by LPR-IV are more noisy and the model is therefore evaluated as an average over years 2 through 6 after IVF treatment.

Table 3. Comparison of long-run child penalty estimates across models

	(1)	(2)		(3)		(4)
	LPR-IV	Event	=	Event-IV	+	Difference
Mother	-0.005	-0.163		-0.030		-0.133
	(0.058)	(0.016)		(0.053)		(0.052)
Partner	0.111	0.012		0.097		-0.085
	(0.065)	(0.023)		(0.067)		(0.070)
Gap (mother - partner)	-0.106	-0.175		-0.127		-0.048
	(0.085)	(0.026)		(0.083)		(0.084)

Note: Table shows estimates of earnings for mother, partner, and the gap. Column (1) shows population average long-run estimates ($p = 2, \ldots, 6$) from the LPR-IV model reported in section 6.2. The long-run estimates for the event and event-model models are evaluated at $a = 6$. Column (2) shows the estimates from the event-model, column (3) shows estimates from the event-IV model, and column (4) 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.

to almost 18 percent in the event-study model. But where the LPR-IV model suggests that none of this gap is driven by mothers, the standard event study in contrast finds large negative and statistically significant effects on maternal earnings, and a very small and nonsignificant estimate for partners.

The estimate for the long-run parental earnings gap from the event-IV specification falls between those of the other two models, at about 13 percent. 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 3 percent. While we cannot reject a zero effect on earnings, we strongly reject the event-study estimates. For partners, the event-IV model estimates an earnings increase of around 13 percent, which is closer to that of the LPR-model, and again different from the null effect in the event-study model.

Table 3 illustrates 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 our event-IV model estimates fertility effects by the age of the child, and their 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

$$\text{Fertility}_{ip} \equiv \sum_{a \geq 0} \mathbb{1}_{\{\text{time since IVF}_i = p\}} \mathbb{1}_{\{\text{age 1st child}_{it} = a\}} \tag{8}$$

Substituting the event-IV first-stages (6) into (8) 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

$$\begin{aligned} y_{it} &= \sum_{a \geq 0} \delta_a \mathbb{1}_{\{\text{age 1st child}_{it} = a\}} + \dots && (9) \\ &= \sum_{a \geq 0} \delta_a \Big(\sum_p \pi_{ap} \mathbb{1}_{\{\text{time since IVF}_i = p\}} \times \text{success}_i \Big) + \dots && (10) \\ &= \sum_p \mathbb{1}_{\{\text{time since IVF}_i = p\}} \sum_{a \geq 0} \delta_a \pi_{ap} \text{success}_i + \dots && (11) \end{aligned}$$

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

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

This shows that the fertility effect γ_p is a weighted average of the child penalties 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 penalties for younger children ($\delta_{a<p}$) are negative.

We report the estimated weights 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 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 penalties for having a child any younger than six years old (i.e. $a < p$) are given a negative weight. As the estimated penalties are negative, this weighting biases the fertility estimates in the LPR-IV model towards zero relative to the contemporaneous child penalty with the positive weight. We show that this pattern holds for all p in appendix figure A10. On the very short run ($p = 0$) the fertility effect γ_0 is equal to the earnings effect δ_0, but with time the

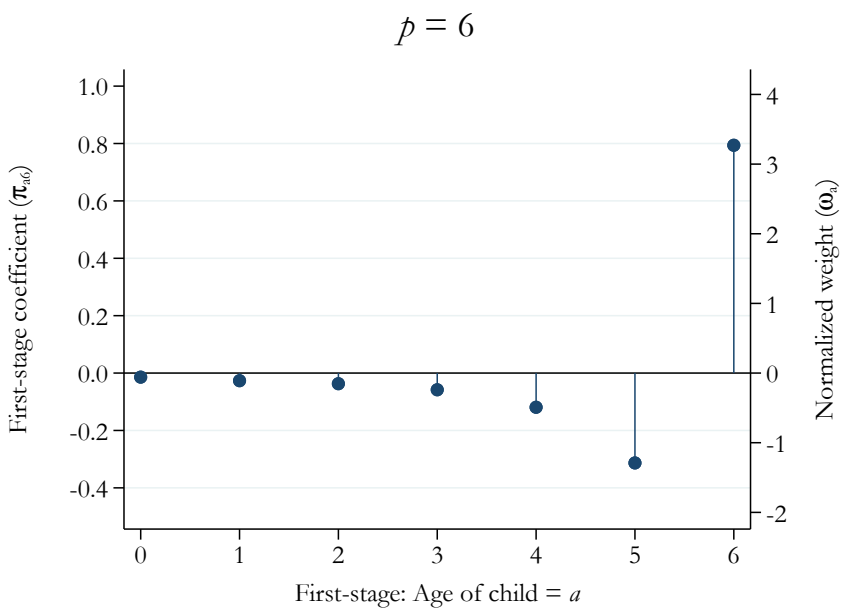


Figure 6. Mapping the first stages of LPR-IV to event-IV

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 six years or younger in the event-IV model.

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 (3) and (4). The mapping is illustrated in appendix figure A11 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.[20]

7.2 Event-IV and Event

The estimates for the earnings effects differ vastly 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

[20]Note that the equivalence requires that all controls are interacted with $\mathbb{1}_{\{\text{time since IVF}_i^1 = p\}}$. Our 2SLS event-study specification in (5) is more parsimonious, which explains why the estimate do not exactly line up, but also shows that this has not consequences for our estimates.

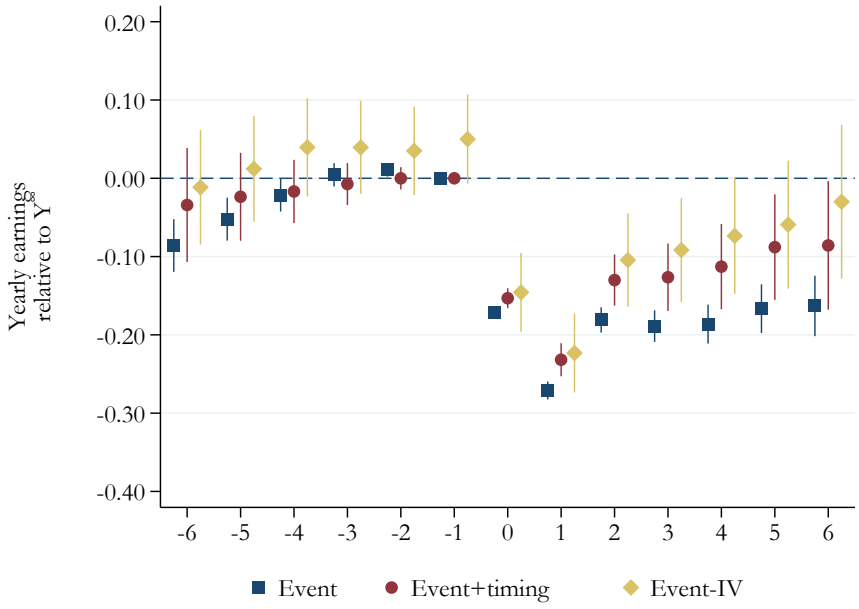


Figure 7. Event vs. event-IV

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.

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

As seen in figure 7, controlling for timing substantially attenuates but does not completely eliminate the pre-trends. Meanwhile, there is a significant reduction in post-birth effects of having a child on earnings. Where the penalty was about 20 percent in the standard event-study setup, controlling for timing more than halves the size of the penalty to about ten percent.

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

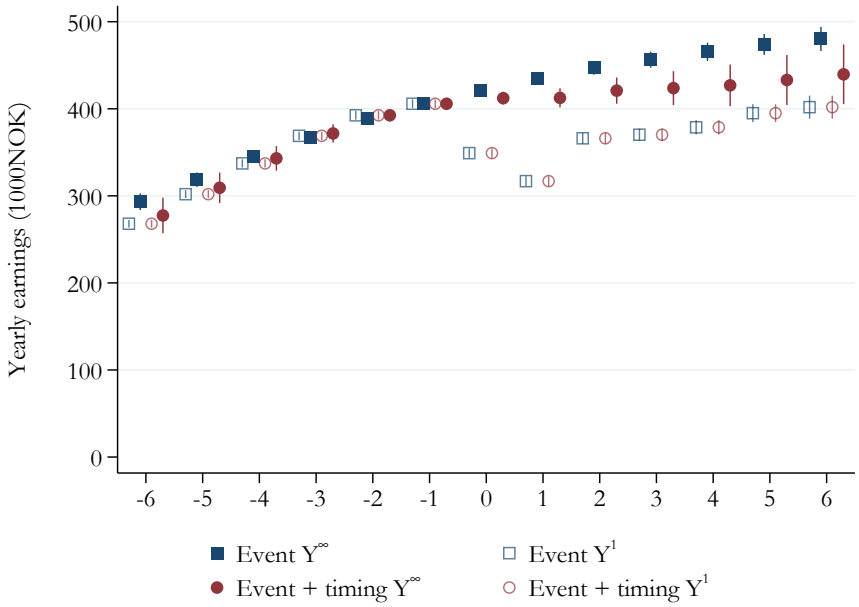


Figure 8. Counterfactual earnings profiles. Event-study.

Note: This figure shows the estimated potential earnings without child (Y^{∞}), and with child (Y^{a}), as estimated from the event-study model. Figure shows estimates with and without controls for time relative to first IVF attempt.

by a sharp drop in the first year after birth. Earnings growth then recovers and after three years women appear to be back on a new age-earnings profile on a lower level, but comparable slope, such that there is a permanent and constant wedge between wages for women with and without children.

The counterfactual earnings profile without a child, Y^{∞}, is slightly 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.

Given these findings one may wonder whether IVF mothers are unusual in that their fertility is planned. Data from the Norwegian Mother and Child Cohort Study (Magnus et al., 2006) show that in the broader population 82 percent of mothers in Norway report that that their child came from a planned pregnancy, and IVF mothers

are therefore typical in this respect.

In a final step we compare the event study estimates that control for timing to the full event-IV estimates. Figure 7 shows that once we control for timing in the event-study model, the fertility effect estimates are much more similar to our event-IV model estimates – to the extent that the differences are no longer statistically significant. This is not surprising: there are by construction no never-takers to our instrument, and had there also been no always takers, that is, if women could not have children without an IVF treatment, then the event-study and event-IV estimates are identical after controlling for the endogenous component of fertility, namely the timing of the fertility attempt. In our application, as much as 80 percent of fertility is channeled through the IVF treatment, which means that the compliers to our instrument are very similar to the population that provides the identifying variation in the event model that adjusts for the timing of the fertility attempt. This is also shown in appendix table A3 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.

7.3 Alternative event-study estimators

Event studies often assess the credibility of the exogeneity or parallel-trend assumption by evaluating the pre-trends. 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 A12 reports event-study estimates that adjust for the baseline of a linear extrapolation of the pre-trend into the post period. The figure shows that the adjusted results exacerbate the bias relative to the standard event-study specification. The reason is that the sign of the selection bias reverses after birth as seen in figure 8, which results in counterfactual earnings estimates that are even higher with extrapolated pre-trends than in the standard event-study.

In the traditional event-study model, both previously treated and untreated observations are used to estimate the counterfactual for a treated unit at any point in time. This is a valid approach only under the assumptions implicit in the model specification of equation (2), such as treatment effect homogeneity and 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., 2022; 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 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 treatment effect heterogeneity recover earnings estimates that are in line with our event-IV model, we apply the imputation estimator of Borusyak et al. (2022) to our sample of IVF women. The Borusyak et al. (2022) approach estimates the counterfactual Y^{∞} on the not-yet-treated observations and includes individual fixed effects which subsume confounding level effects associated with fertility timing. Results are plotted in appendix figure A13 along with the conventional event study estimates.[21] The figure shows that for both mothers and their partners the extrapolation approach estimates even larger negative child penalties, and therefore aggravates the bias relative to the standard event study specification. This is consistent with the results based on the extrapolation of pre-trends. Estimates based on Callaway and Sant'Anna (2021) (not reported here) are very similar.

The major difference between this estimator and the event-study specification where we control for timing (figure 7), is that the latter specification allows earnings profiles to depend on fertility timing rather just the a heterogeneous but time-constant level effect.

8 Conclusion

Social scientists and policy makers have devoted considerable efforts in understanding the drivers of the gender wage gap. In particular, there has been significant attention on how parenthood, specifically motherhood, may 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, while 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.

[21] The not-yet-treated observations do not cover the full sample period. Moreover, adding fixed-effects comes at the cost of having to drop the last year of our data. Taken together this means that we can only estimate the effects up until $t = 5$.

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 women'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, thereby taking remaining endogenous sources of fertility into account, we estimate longer-run earnings effects for mothers of around 4 percent, which is only about one fourth of the effect size uncovered by a standard event-study setup in the same sample. We also find indications of positive earnings effects for partners, whereas the conventional event-study model estimates no effect on partners.

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 is driven purely by negative effects on maternal earnings, while the estimated gap in the event-IV model is driven by the positive effect estimates for partners. 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.

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

Anderson, D. J., Binder, M., and Krause, K. (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 Havnes, T. (2019). Child care, parental labor supply and tax revenue. *Labour Economics*, 61:101762.

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

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

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

Bedard, K. and Deschênes, O. (2005). Sex Preferences, Marital Dissolution, and the Economic Status of Women. *Journal of Human Resources*, XL(2):411–434.

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

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

Borusyak, K., Jaravel, X., and Spiess, J. (2022). Revisiting Event Study Designs: Robust and Efficient Estimation.

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

Callaway, B. and Sant'Anna, P. H. C. (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 D'Haultfœuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, 110(9):2964–96.

Drange, N. and Havnes, T. (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., Joensen, J. S., Johansen, E. R., and Veramendi, G. F. (2022). The labor market returns to delaying pregnancy. Unpublished working paper.

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

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

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

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

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

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

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

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

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

Magnus, P., Irgens, L. M., Haug, K., Nystad, W., Skjærven, R., and Stoltenberg, C. (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.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*.

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

Stevenson, B. and Wolfers, J. (2007). Marriage and Divorce: Changes and their Driving Forces. *Journal of Economic Perspectives*, 21(2):27–52.

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

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

A Appendix For Online Publication

A.1 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

	Failure (1)	Success (2)
Mother characteristics		
Number of IVF attempts	3.31	1.81
Any success	0.43	1.00
Fertility, endpoint	0.65	1.00
Total number of children	0.96	1.53
1 children	0.37	0.53
2 children	0.24	0.42
3 children	0.03	0.05
4 children	0.00	0.00
Age	32.1	31.3
Education		
- Compulsory	0.15	0.12
- High School	0.24	0.23
- Bachelor	0.41	0.44
- Master	0.20	0.21
Earnings (1000 NOK)	362.8	362.6
Hours (FTE)	0.88	0.88
Employed	0.87	0.88
Hourly earnings (NOK)	221.3	221.1
Sickness absence days	15.1	14.7
Visits to general practitioner	2.53	2.47
Visits to general practitioner registered with psychological diagnosis	0.14	0.14
Hospital days	2.21	1.95
Partner characteristics		
Age	35.3	34.5
Female	0.01	0.02
Education		
- Compulsory	0.17	0.16
- High School	0.39	0.37
- Bachelor	0.27	0.29
- Master	0.17	0.18
Earnings (1000 NOK)	455.1	454.4
Hours (FTE)	0.84	0.84
Employed	0.87	0.87
Hourly earnings (NOK)	281.2	283.1
N Women	6 881	3 152

Notes: Table shows descriptive statistics for women who had at least one IVF trial who had at least one IVF trial over the period 2009 to 2016, by success at first trial. 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.

A.3 Comparing event estimates for IVF mothers to non-IVF mothers

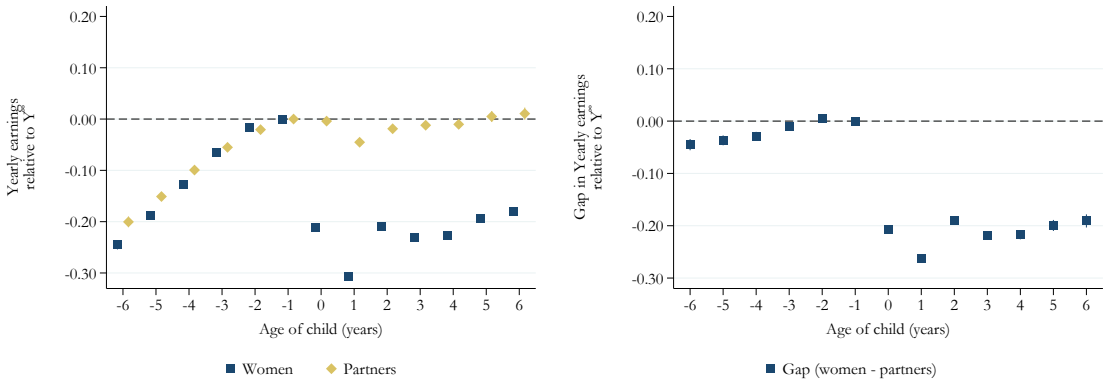
The external validity of our findings would be challenged if mothers conceiving through IVF respond differently to having children than other women. In this section we provide evidence that while there are some observable differences between IVF mothers and non-IVF mothers, they respond very similarly to having children in terms of earnings. We start by reporting the fertility effects using the regular OLS event-study specification (2), estimated on non-IVF mothers, in figure A1.

In figure A1(a) both women and partners display a comparable pre-trend leading up to birth, indicating again that those who have children earlier are on relatively steeper age-earnings profiles compared to those who have children later. These pre-trends are somewhat steeper than what we saw in the IVF-sample (cf. figure 3).

Following birth, earnings changes are almost identical to those found in the IVF-sample: non-IVF mothers see a sharp drop in earnings of about 30 percent which then attenuates somewhat and stabilizes in the neighborhood of 20 percent in the longer run. Partners see almost no change in earnings following childbirth.

Figure A1(b) shows the earnings difference between non-IVF women and their partners. As in the IVF-sample, both parents follow a similar upward-sloping trend in earnings, yielding no discernible pre-trend, but there is still a substantial difference after birth.

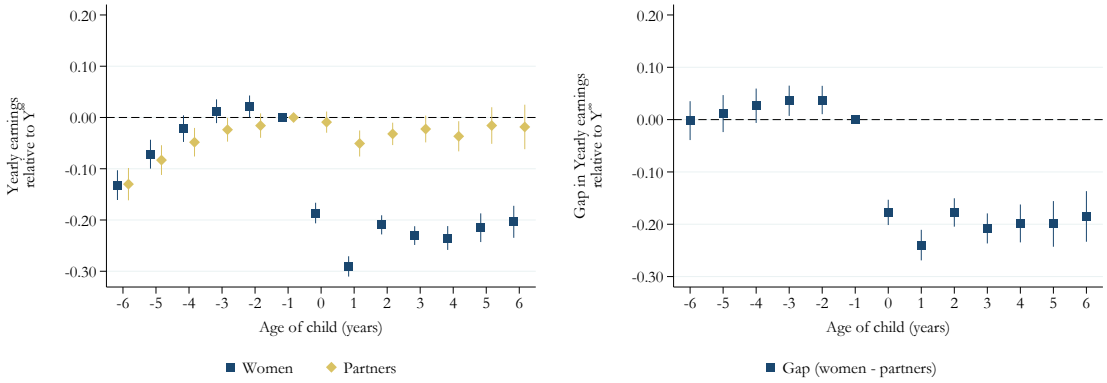
To further investigate whether the compositional differences are important figure A2 reports results for estimates where our sample of IVF women is re-weighted to match the composition of non-IVF mothers in terms of education and age. This exercise yields estimates which are again very similar to those estimated in the sample of non-IVF mothers, suggesting that differences in observable characteristics are not limiting the external validity of these results.



(a) Earnings **(b)** Gap in earnings

Figure A1. Event study estimates of first child on women's earnings. Non-IVF sample.

Note: OLS event study estimates from specification (2). Panel (a) shows effects separately for women and partners, panel (b) shows difference between women and partners. Estimates are scaled relative to each gender's counterfactual earnings (Y^{∞}), as described in section 4.4. Sample is non-IVF mothers who had their first child over the same time period.



(a) Earnings **(b)** Gap in earnings

Figure A2. Event-study estimates of first child on women's earnings. IVF sample reweighted by composition of non-IVF women.

Note: Event study estimates from specification (2), estimated on IVF sample reweighted by the age and education of non-IVF women. Panel (a) shows effects separately for women and partners, panel (b) shows difference between women and partners. Estimates are scaled relative to each gender's counterfactual earnings (Y^{∞}), as described in section 4.4.

A.4 LPR-IV First Stage

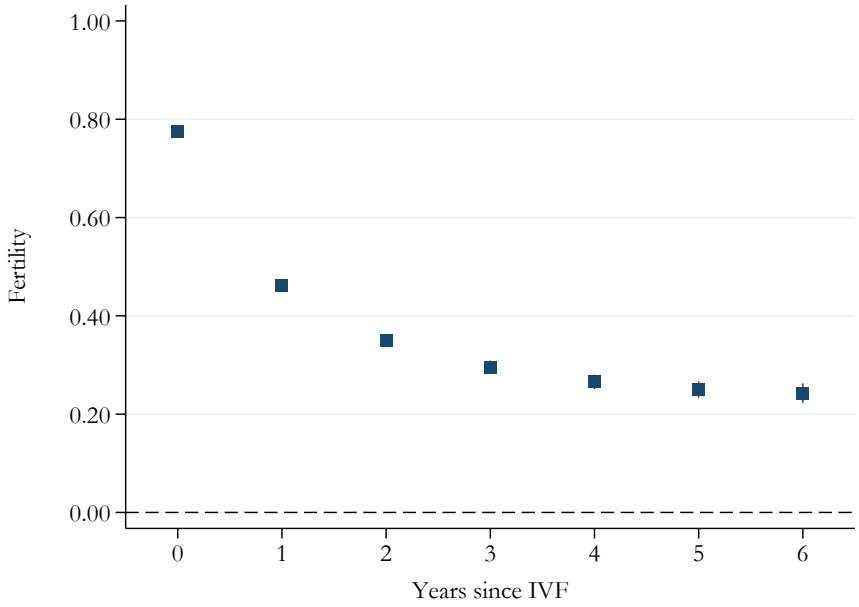


Figure A3. First stage. LPR-IV.

Note: First stage estimates using the IV model of Lundborg et al. (2017) as described in equation (3) 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

In this section we explore the child penalty in terms of a broader set of labor market outcomes. We first disentangle the estimated effect on earnings into hours worked, employment and wage rate. We show these results for the event model (figure A4), the LPR-IV model (figure A5), and the event-IV model (figure A6). For each outcome, we show estimates for mothers and partners in the left-hand column and the difference in outcomes between the two in the right-hand column.

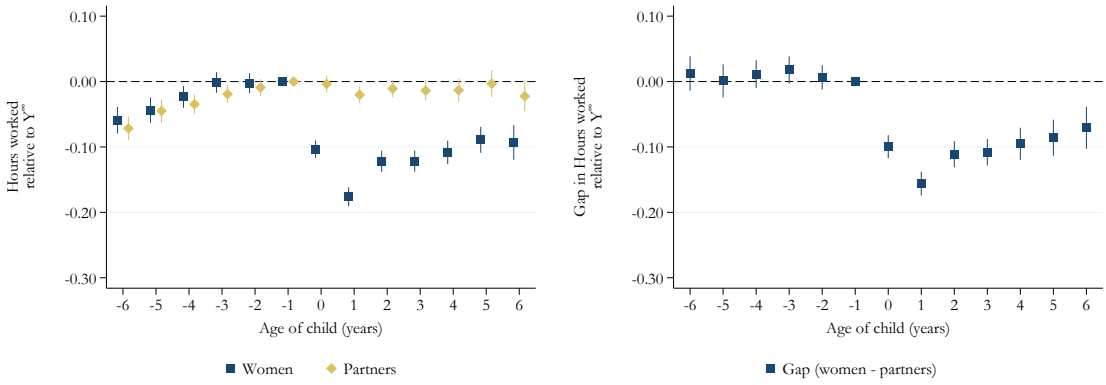
We start by describing results from the event-study models presented in figure A4. These estimates suggest that mothers reduce hours worked by 10 percent in the longer-run (panel (a)) but are only 4 percent more likely to exit the labor market (panel (c)). They receive lower hourly earnings with point estimates suggesting a 10 percent reduction in the longer-run (panel (e)). Partners are unaffected on all these margins, with the potential exception of a minor increase in hourly earnings in the very end of the sample period, such that the estimates for the differences between mothers and partners are driven by effects on mothers (panels (b), (d) and (f)). The event study model therefore suggests that the permanent 18 percent reduction in earnings reflects reduced hours worked and reduced hourly earnings in about equal measure. The effects on wage

rates and hours worked are similar to those found by Kleven et al. (2019) for Denmark, though they find larger effects on employment in their sample.

For the LPR-IV model in figure A5, estimates are noisy but taking point estimates at face value they suggest that there is a negative effect on the difference between mothers and partners along all margins and that the effect is driven by positive effects on partners.

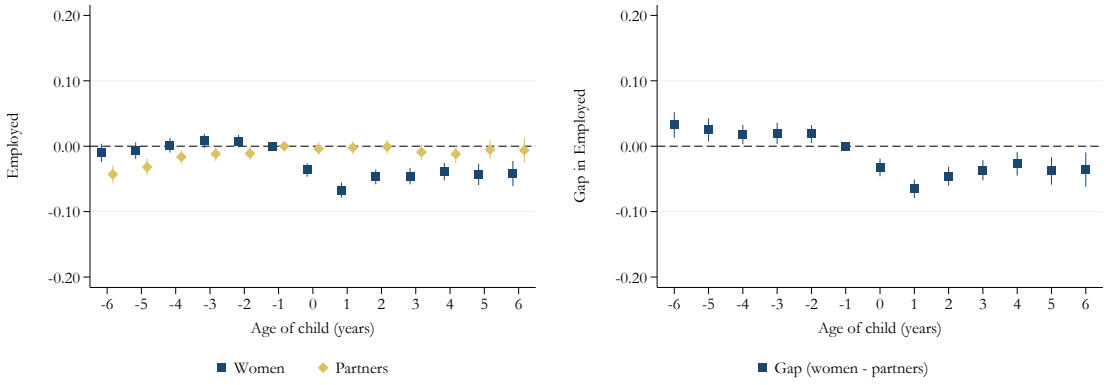
The event-IV results in figure A6 shows that there is a short-run reduction in hours worked for mothers (panel (a)) but this effect diminishes and stabilizes at around 5 percent in the longer-run.[22] Point estimates for partners are about the same magnitude with opposite sign. For employment (panel (c) and (d)), long run estimates are nonsignificant and around 3 percent reduction for mothers, and 4 percent increase for partners. Finally, effects on mothers' long run hourly earnings are close to zero, and again nonsignificant. Our lack of precision makes it difficult to pinpoint the channels that drive the estimated effects on long run earnings, although taken at face value, it seems to be mostly driven by a reduction in hours worked. In comparison, Kleven et al. (2019) find that mothers' earnings reduction is driven by a reduction in both employment, hours and the wage rate, while Lundborg et al. (2017) find effects on hours only in the short run while long run responses are driven by the wage rate.

[22] In comparison, estimated long-run ($t = 6$) reduction in hours worked *given* employment is quantitatively minor, and statistically nonsignificant. These are: -0.057 (0.039) for mothers, 0.045 (0.041) for partners (not shown in paper).



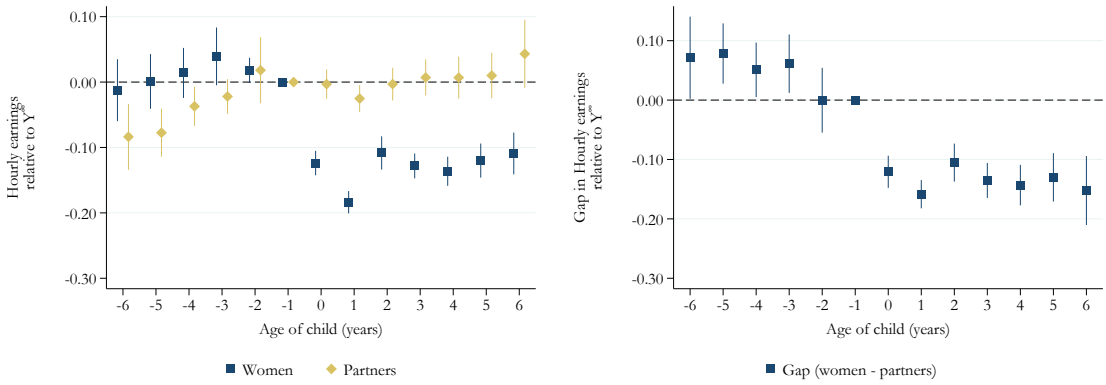
(a) Hours worked

(b) Gap in hours worked



(c) Employment

(d) Gap in employment

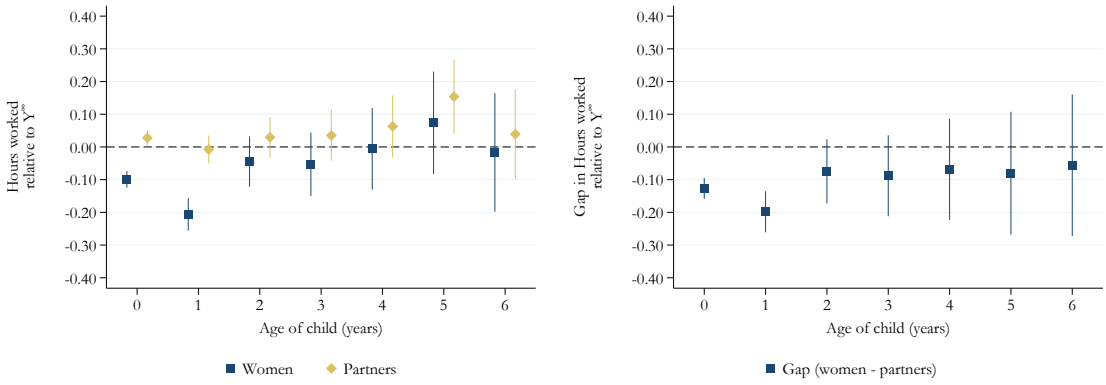


(e) Hourly earnings

(f) Gap in hourly earnings

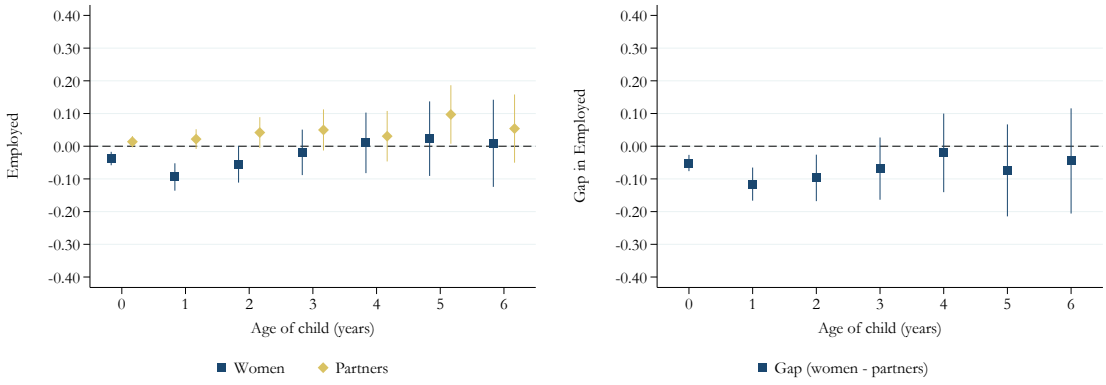
Figure A4. Other labor market outcomes. Event.

Note: Event-study estimates from specification (2). Sample is all women (and their partners) who undergo IVF treatment. 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.



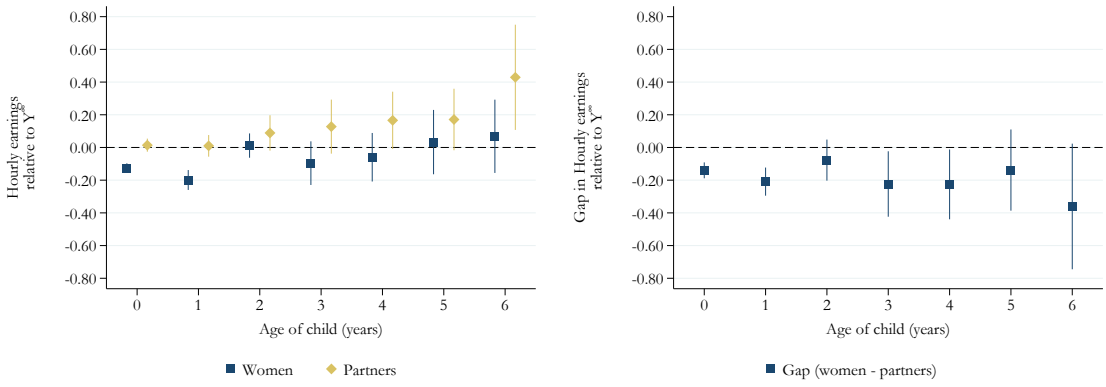
(a) Hours worked

(b) Gap in hours worked



(c) Employment

(d) Gap in employment

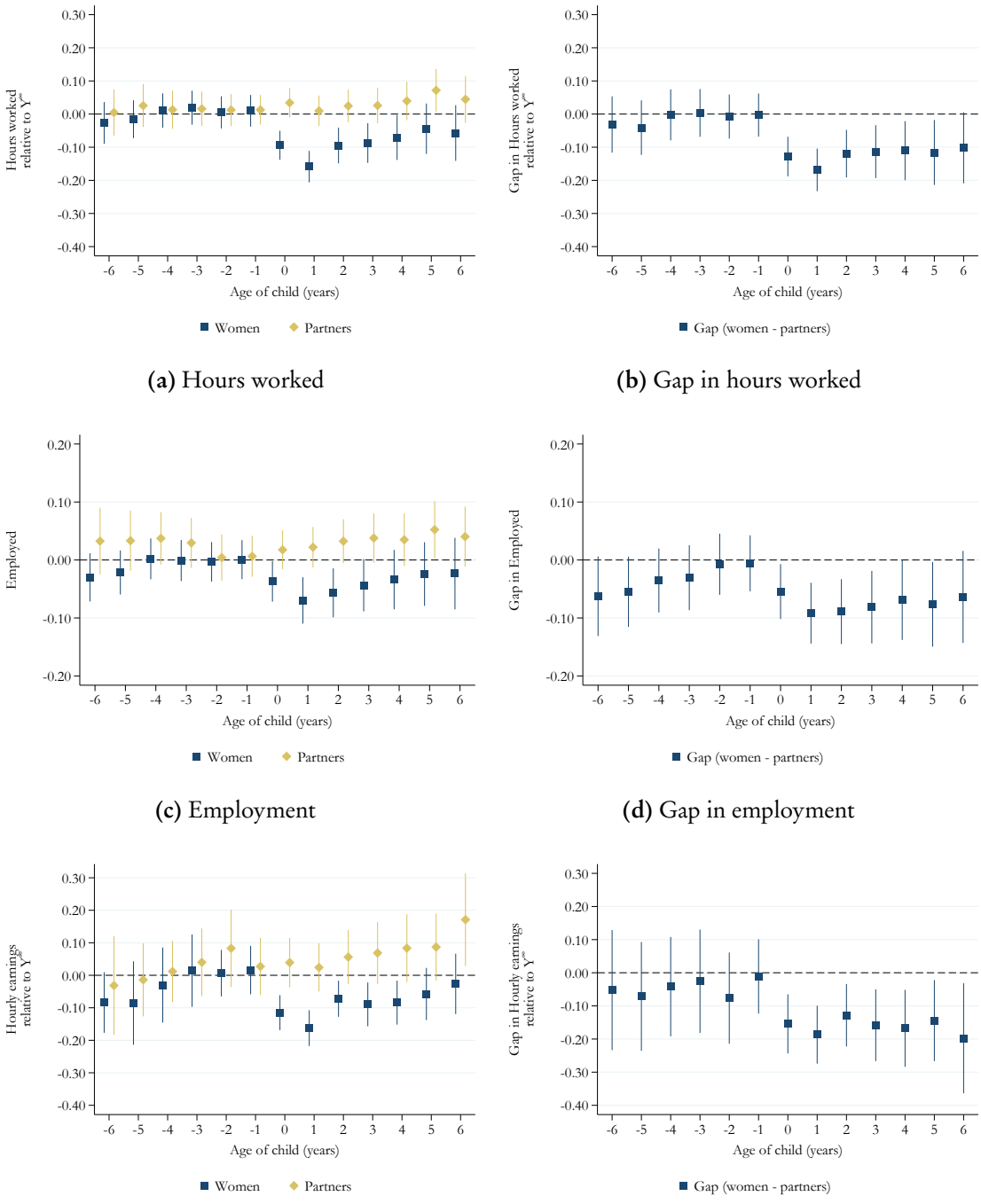


(e) Hourly earnings

(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 (3) on our data. 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

(b) Gap in hours worked

(c) Employment

(d) Gap in employment

(e) Hourly earnings

(f) Gap in hourly earnings

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

Note: Estimated effects of age of child using the event-IV model described in equation (6). 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.

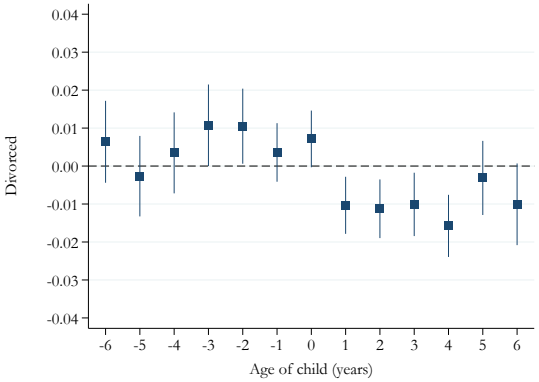
A.6 Non-labor market outcomes

Becoming a parent can have a direct effect on labor market attachment as parents re-optimize time spent at home and work. It is, however, likely that non-labor market outcomes are directly affected as well. For instance, women who succeed in their IVF trial may have different divorce probabilities than those who fail their IVF trial. Divorcees may in turn work more to make up for lost spousal income (Bedard and Deschênes, 2005; Stevenson and Wolfers, 2007). In figure A7a we estimate the effect of having children on being divorced. The results show that having children reduces the probability of being divorced slightly for the first few years after birth.

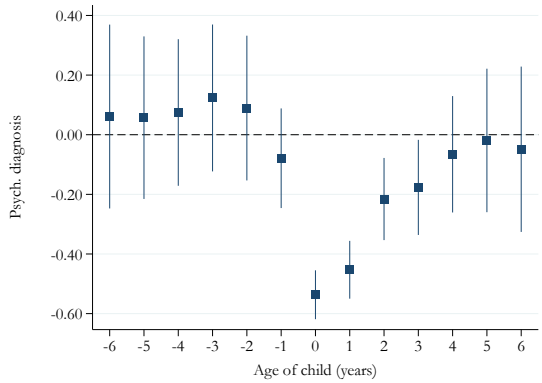
One worry with using IVF as exogenous fertility shocks concerns their potential confounding effects on outcomes if the mental health of women who do not conceive after a trial is adversely affected. Depression might in turn contribute to lost earnings causing a bias towards zero in our estimates. In figure A7b we explore this issue by estimating the effect on the number of visits to the general practitioner that have been registered with a psychological diagnosis. The figure shows a short-lived positive mental health boost from successful trials with a reduced probability of receiving a depression diagnosis. However, in the longer run, e.g. after three years, there are no mental health or other differences between women who have children and those who do not. We also consider the number of visits to the mother's general practitioner following childbirth. Women who successfully conceive after IVF treatments tend to visit their general practitioner more frequently during pregnancy (figure A7c) and in the years following birth.

In figure A8 we investigate the robustness of our estimates to confounding channels. First, we adjust for psychological visits, any visits to the general practitioner and divorce in our estimation as in a mediation analysis. However, the earnings estimates barely move, suggesting that these channels are not confounding our estimates. We interpret this, together with the fact that the share of women adversely affected on this margin is very small, as evidence that this is unlikely to be major source of bias.

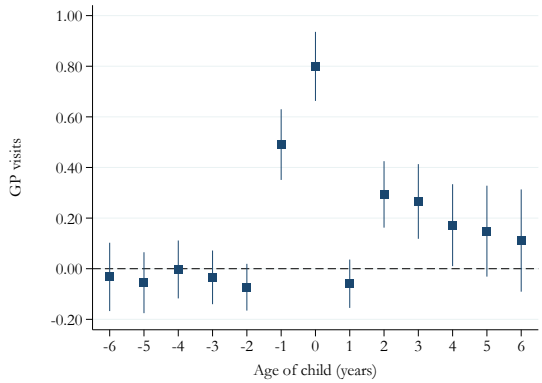
Finally, the estimates above suggested there is a slight imbalance in the years leading up to the first birth. If women who have a child due to a successful IVF earn more than unsuccessful women prior to conception, the estimates of having a child of any age are potentially biased. One way to alleviate this concern is to adjust for average earnings in the years prior to the IVF trial as is done in Lundborg et al. (2017). In figure A8 we also report estimates from specifications that adjust for pre-treatment earnings interacted with pre-treatment education to flexibly control for the potential imbalance in pre-treatment earnings. Results from this specification are shown alongside our baseline results. These findings show that controlling for prior earnings gives very similar child-



(a) Divorce. Event-IV.



(b) Visits to general practitioner registered with psychological diagnosis. Event-IV



(c) Visits to general practitioner

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

Note: Results from our event-IV model shown in equation (6) using visits to general practitioner and psychological diagnoses as outcomes. All estimates are scaled relative to counterfactual earnings (Y^{∞}) as described in section 4.4.

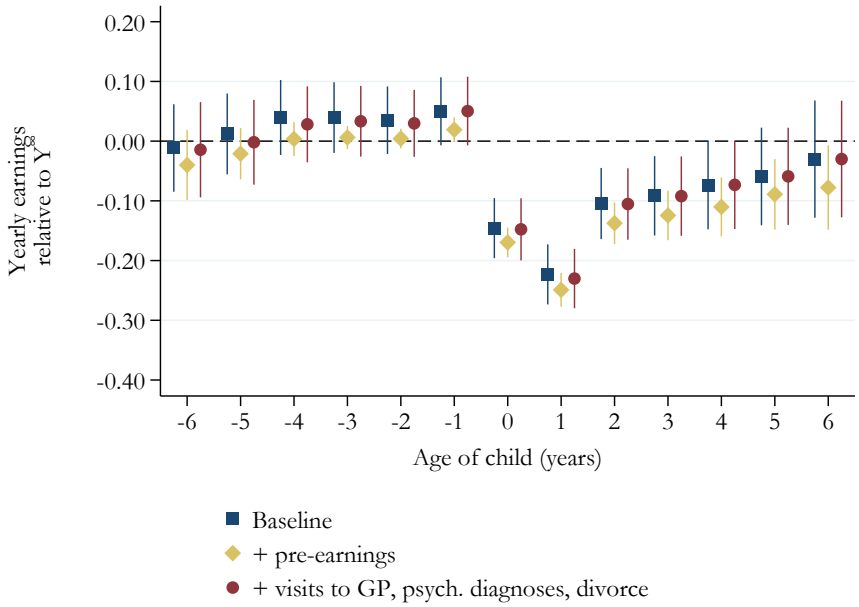


Figure A8. Robustness. Event-IV.

Note: Robustness checks of our event-IV model as specified in equation (6). Figure shows our baseline specification, alongside estimates that include average earnings the four years before the IVF trial interacted with education level, and estimates that include controls for visits to general practitioner and psychological diagnoses. All estimates are scaled relative to counterfactual earnings (Y^{∞}) as described in section 4.4.

penalty estimates, implying that the pre-treatment imbalance in our baseline model is inconsequential for our main results.

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 A9 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.

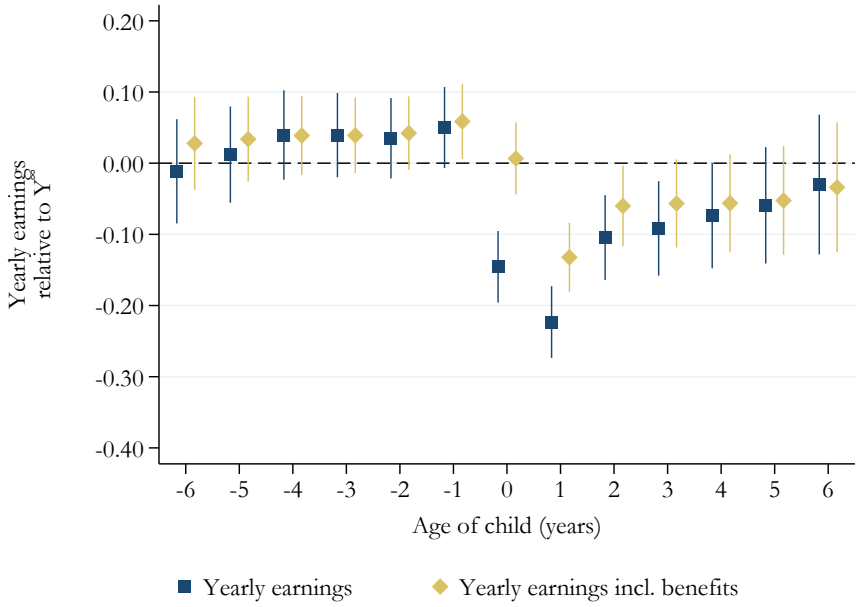


Figure A9. Earnings including benefits. Event-IV.

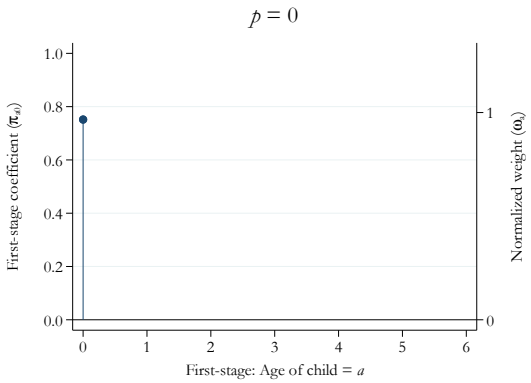
Note: Estimates from our 2SLS event study model as specified in equation (6). Outcomes are earnings and earnings including benefits. Estimates are scaled relative to counterfactual earnings (Y^{∞}) as described in section 4.4.

A.8 Event-IV and LPR-IV

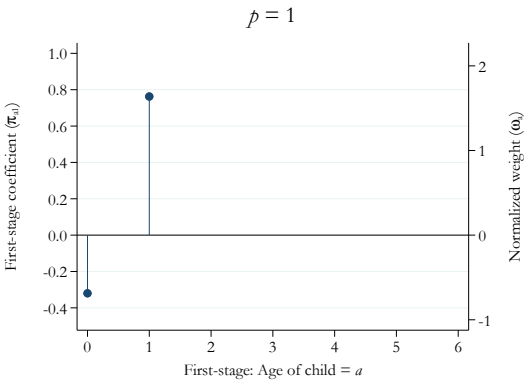
Table A2. First stage F-statistics for LPR-IV and event-IV.

	(1)	(2)
Years since IVF (column 1) / Age of child (column 2)	LPR-IV F-statistic	Event-IV F-statistic
-6		217
-5		255
-4		272
-3		279
-2		301
-1		310
0	247	317
1	239	275
2	222	246
3	199	208
4	174	174
5	148	141
6	118	97

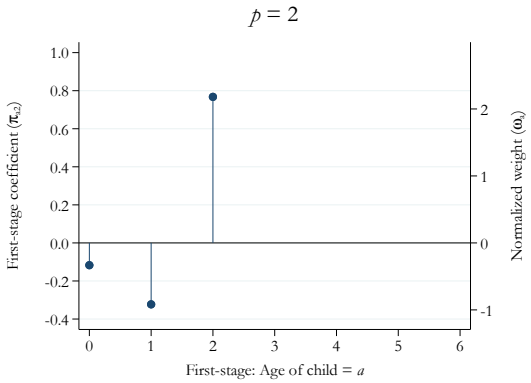
Note: F-statistics for first-stages from the LPR-IV model (equation 4) and the event-IV model (equation 6).



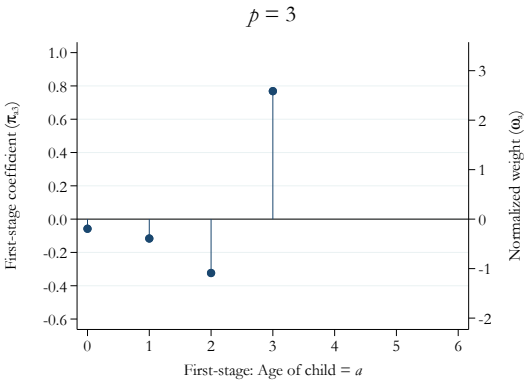
(a) Fertility weight at $p = 0$



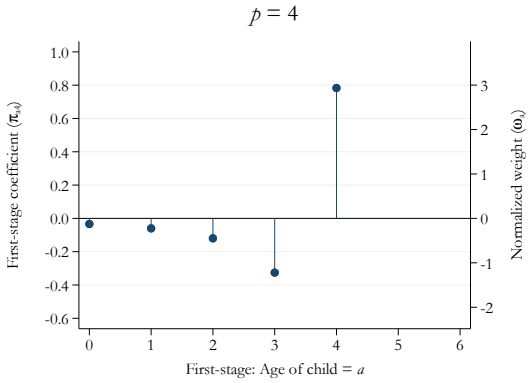
(b) Fertility weight at $p = 1$



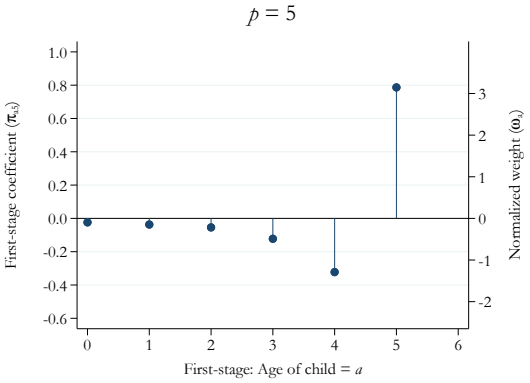
(c) Fertility weight at $p = 2$



(d) Fertility weight at $p = 3$



(e) Fertility weight at $p = 4$



(f) Fertility weight at $p = 5$

Figure A10. Fertility weights

Note: Fertility weights as defined in Section 7.1.

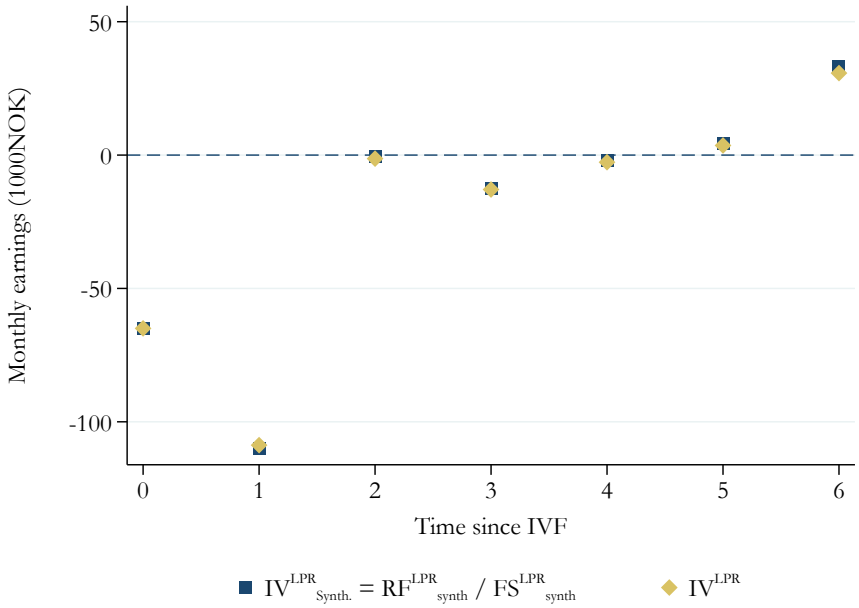


Figure A11. Combining results from LPR-IV and our event-IV model

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 (6). Estimates are scaled relative to counterfactual earnings (Y^{∞}) as described in section 4.4.

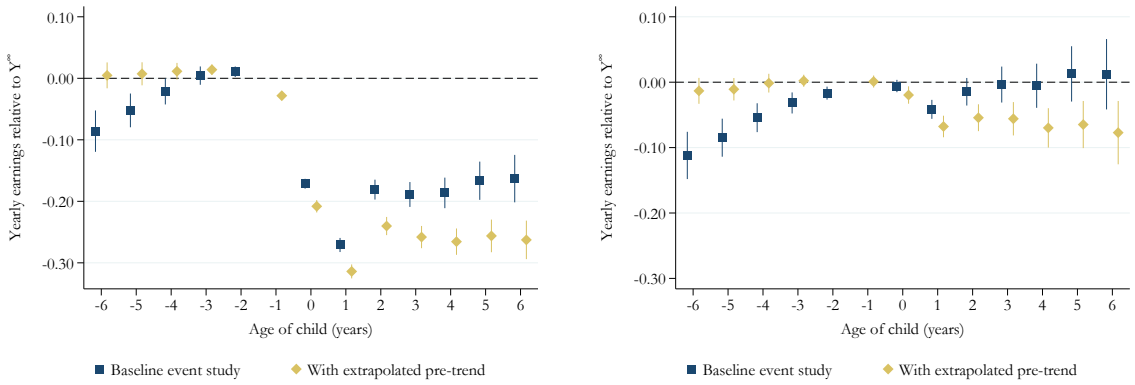
A.9 Complier characteristics

Table A3. Complier characteristics

	All IVF women		Compliers	
	Mean	Std.Dev.	Mean	Std.Dev.
Mother 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 diagnosis	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.

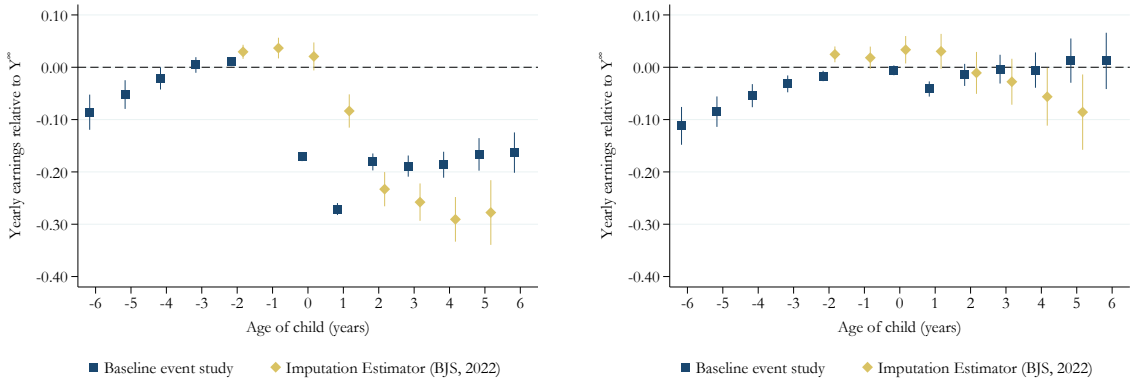
A.10 Alternative event study estimation (cf. Borusyak et al., 2022)



(a) Mother **(b)** Partner

Figure A12. Results based on the extrapolation of the pre-trend

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 that extrapolate the pre-trend,



(a) Mother **(b)** Partner

Figure A13. Results using conventional and imputation estimators

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. (2022) with bootstrapped standard errors.