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

Simon Bensnes[†] Ingrid Huitfeldt[‡] Edwin Leuven[§]

Abstract

This paper reconciles estimates of the labor market effects of having children for women undergoing IVF treatment. We derive an event-IV estimator that places the Lundborg et al. (2017) IV and the standard event study estimator on a common child-age scale. All approaches yield a 15 percent increase in the earnings gap between mothers and their partners, but they differ in incidence. Our event-IV estimates imply that maternal losses account for less than half of the gap, whereas the standard event study attributes the gap to lower maternal earnings. Adjusting for IVF-attempt-timing removes most of this difference.

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

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

*This paper has received funding from the Research Council of Norway (grant #256678 and #326391) and was previously circulated under the title "Event Studies, Endogenous Timing, and the Child Penalty." We thank the editor and two anonymous referees for insightful comments. We also thank Martin Andresen, Jon Fiva, James Heckman, Henrik Kleven, Camille Landaise, Magne Mogstad, Hessel Oosterbeek, and seminar participants at 15th IZA and 2nd IZA/CREST conference on Labor market policy evaluation (2019), OsloMet (2019), Statistics Norway (2019), Norwegian Institute for Public Health (2019), BI Norwegian Business School (2020), Zeuthen Workshop in Copenhagen (2021), University of Chicago (2022), and University of Oslo (2023) for comments. This version April 2026.

[†]Frisch Centre.

[‡]BI Norwegian Business School and Statistics Norway.

[§]University of Oslo and Statistics Norway.

1 Introduction

A large literature studies the labor-market effects of children by comparing earnings trajectories around first birth. Early panel studies, including Korenman and Neumark (1992) and Waldfogel (1997), focused on within-woman changes around motherhood. More recent work uses explicit first-birth event-study designs that trace outcomes in event time (e.g., Kunze and Ejrnæs, 2004; Miller, 2011; Angelov et al., 2016; Kleven et al., 2019). These estimators compare women who give birth at different times after adjusting for age and calendar year. The implicit design relies on no anticipation and a common-levels condition for untreated potential outcomes across timing cohorts, and requires that fertility timing is uncorrelated with untreated potential earnings. A recent alternative approach, due to Lundborg et al. (2017) (LPR), uses first-IVF success as an instrument for fertility. Under the assumption that IVF success is conditionally random given participation, age, and education, this LPR-IV design identifies fertility effects under monotonicity, and exclusion.

The LPR-IV and event study approaches target different fertility effects. LPR-IV centers time at the IVF attempt and estimates the effect of having had a child of any age by a given horizon since treatment. Event studies center time at birth and estimate age-of-child effects. The two approaches also rely on different identifying assumptions. If fertility is delayed after unsuccessful first-IVF attempts, the LPR-IV estimand mixes effects of children of different ages, and the relevant complier group changes across horizons. If fertility timing is correlated with counterfactual earnings profiles, the event-study assumptions fail.

This paper makes three contributions. First, it derives the formal mapping between LPR-IV and age-of-child effects, showing how delayed fertility induces negative weights in the IVF-attempt-centered estimator. Second, it derives an event-IV estimator that places the standard event-study and IVF-based IV on a common child-age scale, and proposes an increasing-horizon stability check for the timing-homogeneity restriction. Third, it compares standard event-study, LPR-IV, and event-IV estimates in a setting where the timing of fertility attempts is observed, using a sample of couples undergoing IVF treatment. The standard event-study and event-IV estimators are directly comparable because both target age-of-child effects, although they rely on different identifying variation: the event study uses observed differences in birth timing across women under the exogenous-timing assumptions of the event-study design, whereas event-IV uses variation induced by IVF success while adjusting for the timing of the attempt.

Using Norwegian administrative data on IVF treatments, family links, and labor-market outcomes, we implement all three estimators in the same IVF setting. Each implies a long-run increase of about 15 percent in the mother-partner earnings gap, often referred to as the *child penalty*, but they differ in incidence. The standard event study attributes most of the gap to lower maternal earnings, LPR-IV attributes a much larger share to higher partner earnings, and event-IV lies in between. This distinction matters because policy conclusions differ depending on whether the child penalty reflects lower maternal earnings or higher partner earnings. In our event-IV estimates, maternal earnings fall by about 7 percent in the longer run, roughly half the decline implied by the standard event study, while partner earnings rise by about 9 percent. For women, the earnings response mainly reflects employment; for partners, it reflects employment as well as hours and hourly wages.

Finally, we trace why the estimates differ. Relative to event-IV, the LPR-IV estimator increasingly loads on younger-child effects and on births delayed beyond the first IVF attempt, because the relevant complier group changes across horizons. Relative to event-IV, the standard event study implies larger maternal earnings losses. Conditioning the event study on observed time since the first IVF attempt substantially reduces the estimated maternal earnings penalty. The estimated counterfactual earnings profiles show that first births among IVF women occur when earnings growth starts to flatten. In this setting, pre-trends are therefore not informative about the direction of the relevant post-birth selection bias, as fertility timing is related to counterfactual earnings growth.

Our results speak to women undergoing IVF treatment. In Norway, about six percent of births follow IVF. IVF women are slightly more educated and conceive at older ages than women in the broader population, but are otherwise similar on observable characteristics. However, women undergoing IVF are actively trying to conceive and may have adjusted their labor market behavior in anticipation. This may limit extrapolation to the broader population. Births following IVF may also differ from other births in ways that matter for external validity. A key concern is disappointment after an unsuccessful IVF attempt (e.g. Gallen et al., 2023; Bögl et al., 2024; Martinenghi and Naghsh-Nejad, 2025). In the IVF design, disappointment from not conceiving and the effects of having a child cannot be separately identified because they are opposite realizations of the same IVF attempt. We document non-labor-market effects on mental-health-related outcomes and divorce following IVF success or failure, but these patterns do not by themselves resolve this concern. Anticipation and disappointment also complicate comparisons between event-study estimates within the IVF sample and event-study estimates from broader samples of women.

A growing literature studies fertility effects using quasi-experimental designs beyond the standard event study. Gallen et al. (2023) study dynamic non-compliance after unsuccessful conception, and Ilciukas (2024) separates parenthood from timing using IUI sequences. The paper also relates to the broader literature on fertility and female labor supply, including dynamic labor-supply models with endogenous fertility and empirical designs using contraceptives, infertility shocks, miscarriages, twins, and same-sex instruments (e.g. Heckman and McCurdy, 1980; Hotz and Miller, 1988; Miller, 2011; Angrist and Evans, 1998).

The remainder of the paper describes the institutional setting and data, presents the three estimators and their connection, reports the main results and reconciliation exercises, discusses the IV assumptions and external validity, and concludes.

2 Institutional setting, data, and sample

Institutional setting

IVF is an assisted reproductive technology that helps couples have children by fertilizing eggs outside the body and transferring the resulting embryo into the uterus. IVF in Norway is regulated by the Biotechnology Law. Women who have unsuccessfully tried to conceive for a year, and who are in a marriage-like relationship are entitled to three publicly funded treatments. A treatment may include multiple embryo insertions from a single egg retrieval; throughout the paper, we define a trial or attempt as an embryo insertion, which is the margin observed in the reimbursement data. During our sample period, the standard procedure was single-embryo transfer, so IVF generated very few multiple births (Bhalotra et al., 2019; Bhalotra and Clarke, 2019) (in our data, <0.01).

Public treatment is heavily subsidized, whereas private IVF is substantially more expensive. The copayment for the first three treatments at a public hospital is about NOK 6 000 (USD 670 in 2019) per treatment, and covers medicines and pharmaceutical expenses. Private institutions offer an alternative to public hospitals and comprise 15-20% of the market. Private options are considerably more expensive – around NOK 100 000 (USD 10,900) for a single treatment – but may have shorter wait times and more flexibility in terms of age requirements.

Norwegian labor-market institutions may attenuate earnings losses from children. During our study period, parents were entitled to about one year of parental leave, either at full replacement for a slightly shorter period or at 80 percent replacement for ten weeks longer.

Data, variables, and sample

The empirical analysis combines administrative registers from Statistics Norway and the Norwegian Directorate of Health. Unique personal identifiers allow us to link health records to population registers for the full Norwegian population, including family links to partners and children. These data are available through 2022.

Public-hospital IVF treatments are observed in the Norwegian Patient Registry from 2008 to 2017. We identify IVF attempts using procedure code “LCA 30 - Transfer of zygote or embryo to uterus in assisted fertilization” and also construct annual hospital days from these records. We also use primary-care records from KUHR for 2006 to 2017. From these data we construct annual GP visits, an indicator for at least one visit coded with a psychological symptom or disorder, and a separate indicator restricted to severe diagnoses (Appendix Table A1).

Our labor-market outcomes come from the employer-employee registry, available from 2004 to 2022. We study four measures: yearly labor earnings excluding parental leave benefits (our main outcome, in 2015 NOK), employment, contracted hours, and hourly earnings. Employment equals one if annual labor income exceeds the substantial gainful activity threshold.[1] In the main text we focus on earnings and report the other outcomes, as well as earnings including social benefits, in the appendix.

Our main sample consists of 10,033 women whose first observed IVF trial occurred between 2009 and 2016 and who had no children beforehand. We exclude IVF observations in 2008, the first year IVF can be identified, to better isolate first-time treatment. We further restrict the sample to women aged 18 or older who were registered with a partner in the year of the first IVF treatment.[2] We link each woman to the partner registered at the time of her first IVF attempt and hold this link fixed throughout. Partner outcomes — such as labor supply and earnings — are thus attributed to this partner regardless of later union dissolution or childbearing with a different partner. For comparison, we construct a non-IVF sample of women who had their first child in the same period as the successful IVF women and were registered with a partner in the year of conception.

Table 1 reports descriptive statistics, with the full version in Appendix Table A2. Following Lundborg et al. (2017), we define an attempt as successful if the woman gives birth within five to ten months of the trial and no other trial occurs in between. Women undergo 2.8 IVF attempts on average, the first-trial success rate is 31 percent, and 83 percent eventually have at least one child. Relative to non-IVF mothers, IVF women are older, only slightly more educated, and have higher pre-trial earnings and labor supply. The same pattern holds for partners, while health differences are modest.

[1]The substantial gainful activity level ("basic amount") was equivalent to NOK 90,068 in 2015.

[2]Only women in stable unions are eligible for public IVF treatment. Because stable unions need not be formal marriages, some partnerships are not observed in the administrative data. Restricting to registered partners excludes 14 percent of IVF women and 46 percent of non-IVF women.

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

	IVF (1)	Non-IVF (2)	Difference (3)	
Woman characteristics				
Number of IVF attempts	2.84			
Success, first trial	0.31			
Success, endpoint	0.63			
Fertility, endpoint	0.83	1.00	-0.17	(0.00)
Total number of children	1.47	1.97	-0.50	(0.01)
Age	31.82	28.40	3.41	(0.05)
Education				
- Compulsory	0.14	0.17	-0.03	(0.00)
- High School	0.24	0.23	0.01	(0.00)
- Bachelor	0.42	0.41	0.01	(0.01)
- Master	0.20	0.19	0.01	(0.00)
Earnings (1000 NOK)	362.72	285.54	77.18	(1.86)
Hours (FTE)	0.88	0.78	0.10	(0.00)
Employed	0.80	0.67	0.14	(0.00)
N Women	10 033	109 791		

Notes: Column (1) shows descriptive statistics for women who had at least one IVF trial over the period 2009 to 2016. Column (2) shows descriptive statistics for women who had their first child without IVF treatment during the period 2009 to 2017. By construction, this includes only women who have at least one child. Column (3) shows the difference and corresponding standard error. Labor market outcomes are measured as averages over the four years prior to the first IVF trial, or, for non-IVF women, prior to the approximate conception date. Education is measured in the calendar year before the IVF attempt / approximate conception date. Age is defined as the maternal age at the date of the IVF attempt / approximate conception date. The full table, including partner and health characteristics, is reported in Appendix Table A2, additional descriptive statistics on fertility outcomes in Appendix Table A3. Descriptive statistics when non-IVF women are reweighted to match IVF women are reported in Appendix Table A4.

3 Estimating the effects of fertility on labor market outcomes

This section presents the empirical framework. We first introduce the standard birth-centered event study and the IVF-based IV approach of Lundborg et al. (2017) (LPR-IV). We then connect the two by showing how the LPR-IV estimand can be written as a mixture of age-of-child effects, which motivates our event-IV estimator. Finally, we describe how we scale effects and define the child penalty.

3.1 Event study

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

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

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

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

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

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

This specification anchors the counterfactual wage profile to a year prior to birth ($a = -1$). The coefficients on the child-age dummies, δ_a, allow for age-of-child specific effects on the outcome y_{it}. The counterfactual wage profile consists of controls x_{it} which adjust flexibly for women's age using dummy variables, and calendar year dummies τ_t. The main outcome and summary measure of women's labor supply that we consider is yearly earnings from employment, but we also look at additional outcomes such as hours, hourly wages and employment.

Equation (2) is the pooled event-study specification that we use throughout the paper. When the estimation restricts attention to mothers, the counterfactual outcome profile (and therefore δ_a) is identified from pre-birth (untreated) outcomes, coming from differential timing of first birth across women.

In our application, the baseline event-study is estimated on IVF women who eventually have a first child, so the relevant comparison is across timing cohorts within the IVF sample rather than between IVF women and women who may never attempt conception.

To relate the pooled event-study coefficient δ_a to underlying cohort-specific causal effects, let the cohort-specific average treatment effect (ATT) at event time a for timing-cohort $T_i = g$ be

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

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

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

then δ_{ag} can be identified using not-yet-treated units.[3]

In the event-study specification (2), which imposes a common set of event-time coefficients across timing cohorts, the OLS estimates of δ_a equal the simple ATT at event time a only under a cohort-homogeneity restriction $\delta_{ag} = \delta_a$ for all g; without homogeneity, the estimates of δ_a are (possibly non-convex) weighted averages of the δ_{ag}.

3.2 Fertility effects of IVF (LPR-IV)

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

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

Let IVF_i be the calendar year of woman i's first IVF treatment and re-index time by years since the first IVF treatment: $p = t - IVF_i \in \{0, 1, 2, \ldots\}$, so $y_{ip} \equiv y_{i,t=IVF_i+p}$, etc. Define the instrument $\text{success}_i \in \{0, 1\}$, equal to one if the first IVF treatment leads to a birth roughly nine months later.[4] The "treatment" at horizon p is the extensive-margin indicator

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

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

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

where γ_p is the extensive-margin fertility effect of interest. To account for the endogeneity of fertility they instrument Fertility_{ip} with first-IVF success_i:

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

Here x_{ip} includes flexible dummies for (woman's) age, and both equations control for IVF year, IVF_i (i.e. η_{IVF_i} and ζ_{IVF_i}) , and time since first IVF treatment, p (θ_p and ξ_p). In

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

[4]Our focus on first-IVF success abstracts from repeated treatment assignment. For related discussion of repeated-treatment settings and the IVF context, see Ketel et al. (2024) and Ilciukas (2024).

addition all controls are interacted with education to support conditional independence.[5] We refer to the 2SLS estimate of γ_p as the LPR-IV estimator.

Identification of γ_p at horizon p relies on: (i) instrument relevance ($\pi_p \neq 0$), (ii) conditional independence of success$_i$ given age, education, and IVF_i, (iii) monotonicity (mechanical here in the sense that first-IVF success weakly increases fertility at every horizon), and (iv) an exclusion restriction stating that, for $p \geq 1$, IVF success affects y_{ip} only through Fertility$_{ip}$. Under these conditions, the 2SLS coefficient $\hat{\gamma}_p$ has the usual Wald interpretation as an estimate of the local average treatment effect of fertility for women who have a child p years after the IVF attempt if the IVF is successful but not otherwise.

Compliance is horizon-specific:

$$C_p \equiv \{\text{Fertility}_{ip}(1) = 1,\ \text{Fertility}_{ip}(0) = 0\}.$$

A woman can be a complier at short horizons and an always-taker at longer horizons if she gives birth after an unsuccessful first IVF. Lundborg et al. (2017) note that the standard extensive-margin interpretation breaks down if IVF success affects not only whether a woman has had a child by horizon p, but also the age of that child. Figure 1 shows that delayed fertility is empirically relevant: the first stage is largest–close to 80 percent–in the year of the IVF treatment ($p = 0$), and declines thereafter as more women become always-takers.[6] In Appendix A.1, we normalize this short-run birth response to the first post-IVF horizon. As a result, γ_p mixes extensive-margin and timing (age-of-child) effects, and can understate short-run effects when impacts decline with child age. The next subsection shows that γ_p is a first-stage-weighted average of child-age-specific effects.

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

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

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

[5]While Lundborg et al. (2017) estimate (4) separately for each horizon p, we instead pool horizons and impose that the coefficients on the age/education controls and IVF-year effects are common across p, while allowing γ_p, π_p, and the horizon effects θ_p and ξ_p to vary with p. This yields a more parsimonious mapping as shown below. This does not matter for the results. Figure A1 compares the p-stratified estimates with those from the specification in (4), and shows that these are nearly identical.

[6]In the data, time-since-IVF is constructed from the month of the IVF attempt, shifted back by nine months and then grouped into one-year bins. This explains the small positive fertility rate at $p = -1$. Women who are unsuccessful in their first trial may still have children within the first year of treatment.

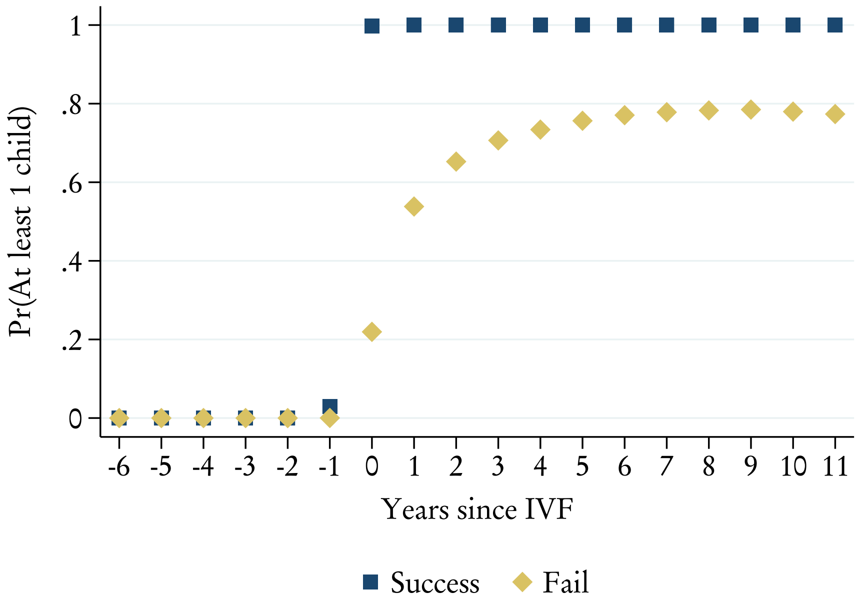


Figure 1. Fertility by success at first IVF trial

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

Substituting these age-horizon first stages into the following identity[7]

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

gives the first-stage (5) of LPR above where the first-stage coefficient π_p maps to the age-horizon specific first-stage coefficients as follows[8]

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

Similarly, write the horizon-p outcome as

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

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

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

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

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

where the tilded terms collect the corresponding coefficients and residual components after substitution.[9] The reduced-form coefficient ρ_p equals

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

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

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

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

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

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

$$\delta_{ap} = \begin{cases} E\left[y^a_{ip} - y^\infty_{ip}\right], & a = p, \\ E\left[y^a_{ip} - y^\infty_{ip} \mid P_i(0) = p - a\right], & a < p, \end{cases}$$

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

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

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

[9] $\tilde{\phi} = \sum_{a=0}^{p} \delta_{ap}\lambda_a + \psi$, $\tilde{\eta}_{IVF_i} = \sum_{a=0}^{p} \delta_{ap}\zeta_{a,IVF_i} + \eta_{IVF_i}$, $\tilde{\xi}_p = \sum_{a=0}^{p} \delta_{ap}\xi_{ap} + \theta_p$, and $\tilde{u}_{ip} = \sum_{a=0}^{p} \delta_{ap}\omega_{aip} + u_{ip}$ are the composite reduced-form objects obtained by substituting the age-horizon first stages into the outcome equation; see Appendix A.1 for the full derivation.

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

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

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

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

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

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

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

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

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

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

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

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

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

This assumption mirrors the cohort-homogeneity condition in the event-study design, corresponds to the 'cohort-invariant dynamic effect' assumption in Ferman and Tecchio (2025) and Angrist et al. (2025), and is closely related to the recursive identification argument in Ketel et al. (2012) and Gallen et al. (2023).

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

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

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

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

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

age a):

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

Adjusting for the time–since–IVF dummies $P_{it,p}$ ensures that each observation is compared only to observations at the same event time (the conditioning set for independence), and interacting success_i with $P_{it,p}$ creates one instrument per p, such that the first stages in (11, FS, event-IV) reproduce the *same* π_{ap} used above. Finally note that, with the full set of time–since–IVF dummies $P_{it,p}$ and because $t = IVF_i + p$, including IVF_i fixed effects is equivalent to including calendar–time fixed effects τ_t; in practice we use τ_t (together with $P_{it,p}$) rather than IVF_i to align with the event-study specification.[11] Although the event-IV equations are written in calendar time t, the relevant compliance variation remains indexed by time since the first IVF treatment, p, through the interactions $\text{success}_i P_{it,p}$, which reproduce the same horizon-specific first stages π_{ap}.

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

The identification arguments above imply that the *levels* of the relevant potential outcomes can also be recovered for the same local populations as the age-specific event-IV effects. In particular, Appendix A.1 shows that the no-child counterfactual level corresponding to age a can be recovered using the transformed dependent variable

$$\tilde{y}_{it}^{\infty,a} \equiv -y_{it} \mathbf{1}\{a_{it} \neq a\}.$$

Intuitively, by negating outcomes in all states except the target age-a state, this transformed regression reverses the original comparison: the age-a coefficient no longer captures the treated-minus-untreated gap, but instead loads on the untreated no-child level. Under timing-homogeneity and counterfactual-level homogeneity, the coefficient on the age-a indicator now recovers the mean untreated counterfactual level for the same local population as the age-specific event-IV estimate.[12] This thus allows us to scale the estimates of δ_a

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

[12] The appendix also shows that the mean treated level can be recovered analogously using a suitable transformed outcome.

relative to the corresponding untreated counterfactual mean.

Finally, the above suggests a simple check of the timing–homogeneity restrictions used in equation (10, event-IV). To identify δ_0 we use only $p = 0$ (no homogeneity needed). To identify δ_1 we add a single homogeneity assumption across the adjacent timing cohorts $p = 0$ and $p = 1$. Moving to δ_2 adds another two homogeneity assumptions: require that the δ_0 effect is invariant between $p = 0$ and $p = 2$ and that δ_1 is invariant between $p = 1$ and $p = 2$. In general, identifying $\{\delta_0, \dots, \delta_P\}$ using information from horizons $p = 0, 1, \dots, P$ requires $P(P + 1)/2$ homogeneity assumptions. As P grows, these assumptions link complier groups that are progressively farther apart in IVF-induced timing. This delivers overidentifying restrictions: each δ_a can be estimated at multiple horizons $P \geq a$ and should be stable if timing–homogeneity holds. We implement this increasing-horizon stability check by re-estimating the age profile for $P = 0, 1, \dots, 11$. Figure A2 shows the age-specific effects are essentially unchanged as the horizon moves from 0 to 11, providing strong support for timing–homogeneity.

To summarize, *i)* centering time on birth renders the treatment invariant to dynamic extensive margin fertility responses over time, *ii)* adjusting for timing through $P_{it,p}$ accounts for the dynamic selection into the fertility attempt, and *iii)* the instrumentation addresses potential remaining unobserved variable bias due to other sources of fertility.[13]

3.5 *Definitions of fertility effects and the child penalty*

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

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

where $y_a^{\infty,women}$ denotes the average counterfactual outcome for mothers of child age a. In the event-study specification, this counterfactual is obtained by subtracting the estimated fertility effects δ_a from the observed outcomes of parents with children of age a. In the IV specification, we follow Abadie (2003) to recover the corresponding counterfactual outcomes.[14] We construct the same measures analogously for partners.

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

[14]In Appendix Table A5 we show complier characteristics using κ weighting.

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

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

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

4 IV assumptions and diagnostic evidence

This section discusses the identifying assumptions behind using success in the first IVF treatment as an instrument and presents diagnostic evidence based on observed predetermined characteristics and pre-treatment outcomes. These patterns are informative about the plausibility of the maintained assumptions, but they do not establish them. Table 2 reports estimates from a regression of pre-IVF earnings (column 1), and IVF success (column 2), on a number of observable predetermined characteristics capturing women's demographics, labor market attachment and health.[15] As in the 2SLS specification in Equation (10, event-IV) and (11, FS, event-IV), all regressions include controls for calendar time, time since IVF treatment, maternal age, and education, which are known predictors of success (CDC, 2012; Groes et al., 2017). Our results therefore rely on conditional exogeneity of success, and not on an assumption that success is unconditionally random.

In column (1), the regression of pre-IVF earnings on background characteristics highlights potential confounders of our instrument. Many of these characteristics are strongly correlated with earnings (our main labor supply measure): women with poorer health, as measured by visits to their primary care physicians, and being diagnosed with a severe 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.

A necessary condition for conditional exogeneity is that IVF success is not correlated with all observable characteristics that predict earnings. Column (2) shows that characteristics predictive of pre-randomization earnings are generally not predictive of the instrument. A joint significance test across all variables is insignificant, yielding a p-value of 0.22.

These results are consistent with exogeneity of IVF success conditional on women's age and education, but they do not establish it. Any remaining confounder would have to be correlated with potential earnings while leaving pre-earnings and pre-trends largely unaffected in the years prior to the IVF attempt.

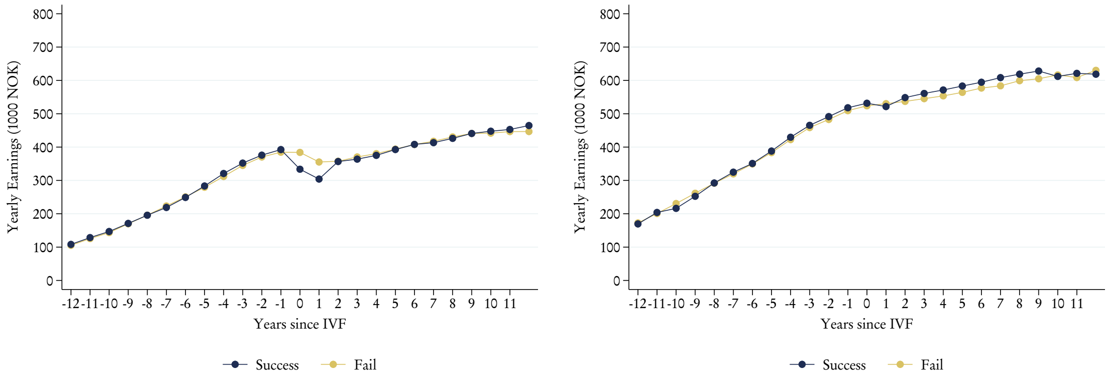
Although Table 2 indicates that any imbalance is likely to be minor, this test is based on an average over the four years preceding the first IVF trial. To make sure that this

[15] In appendix Table A6 we also show the raw means by success at first trial.

Table 2. Predetermined characteristics and IVF success

	Pre-IVF Earnings (100K NOK) (1)		IVF Success (2)	
	est.	*s.e.*	*est.*	*s.e.*
Woman characteristics				
Earnings (100K)			0.004	(0.003)
Hours (FTE)			-0.006	(0.012)
Sickness absence days (/10)	-0.009	(0.002)	0.001	(0.001)
GP visits	-0.034	(0.005)	-0.002	(0.002)
Any (mild/severe) psychological diagnosis	0.027	(0.039)	-0.003	(0.011)
Severe psychological diagnosis	-0.381	(0.056)	0.021	(0.018)
Hospital days (/10)	0.000	(0.002)	-0.001	(0.001)
Partner characteristics				
Age (/10)	-0.223	(0.031)	-0.003	(0.010)
Earnings (100K)	0.115	(0.008)	0.001	(0.002)
Hours (FTE)	-0.269	(0.045)	-0.007	(0.012)
Education (ref. master)				
- Compulsory	-0.143	(0.057)	-0.021	(0.018)
- High School	-0.152	(0.053)	-0.030	(0.015)
- Bachelor	-0.054	(0.054)	0.001	(0.015)
Constant	3.477	(0.248)	0.361	(0.094)
Mean dependent variable	3.38		0.32	
Joint F [p-value]	43.8 [<.001]		1.3 [0.22]	
N Women	10 033		10 033	

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



(a) Women

(b) Partners

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

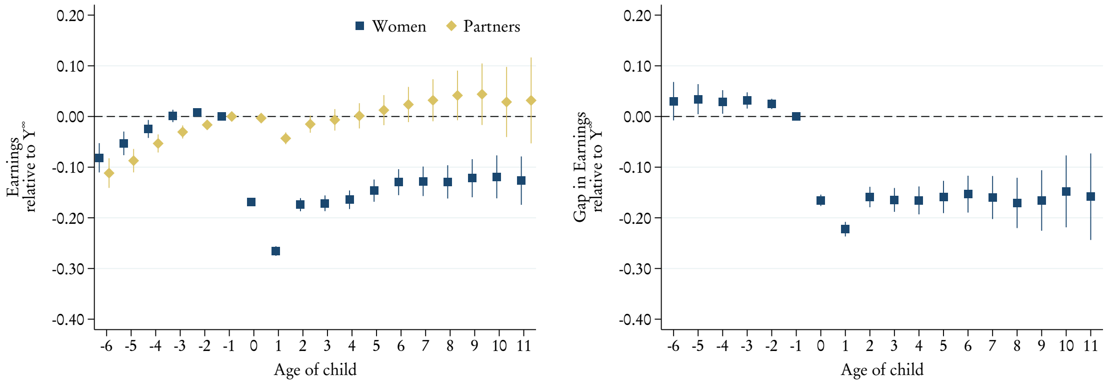
Note: Estimates are regression-adjusted for calendar year, maternal age, and education. Predicted earnings by IVF success and time are averaged over the covariate distribution in the estimation sample. The sample includes all women (and their partners) who had their first IVF treatment between 2009 and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

average does not hide any imbalance in *trend*s, Figure 2 plots average earnings for each year since the first IVF trial, by success, completely adjusting for the controls included in our main specification: calendar time, maternal age, and maternal education. More precisely, we first construct the estimates in the figure stratified by calendar year, maternal education and maternal age. We then compute the population level estimates by averaging across cells for each year since the first IVF trial. We see that to the extent that there is an imbalance it is constant over time and trends in earnings are essentially identical in the 12-year period leading up to the trial. These results are consistent with the maintained conditional-independence assumption for IVF success.

For the exclusion restriction to hold, we require that IVF success affects no variables other than fertility directly. Some argue that disappointment and related outcomes constitute a violation of exclusion (e.g. Gallen et al., 2023). While women who fail to conceive following IVF may experience disappointment, depression, or divorce (e.g. Bögl et al., 2024; Martinenghi and Naghsh-Nejad, 2025), one interpretation is that such outcomes operate through fertility rather than as a direct effect of the trial. We discuss this in detail, along with issues of external validity, in Section 7.

5 Children and labor market outcomes

We now present the estimated effects on earnings for the three different models described in Section 3: 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



(a) Women and partners

(b) Gap

Figure 3. Earnings – Event study estimates

Note: OLS event study estimates from specification (2). Panel (a) shows effects separately for women and partners, panel (b) shows the difference between women and partners. Estimates are scaled relative to each gender's counterfactual earnings (Y^{∞}), as described in Section 3.5. Point estimates are presented in table form in Table A7. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were registered with a partner at the time, and eventually had at least one child (observations = 145,571; women = 8,349). Results for other labor market outcomes are reported in appendix Figures A3 and A4.

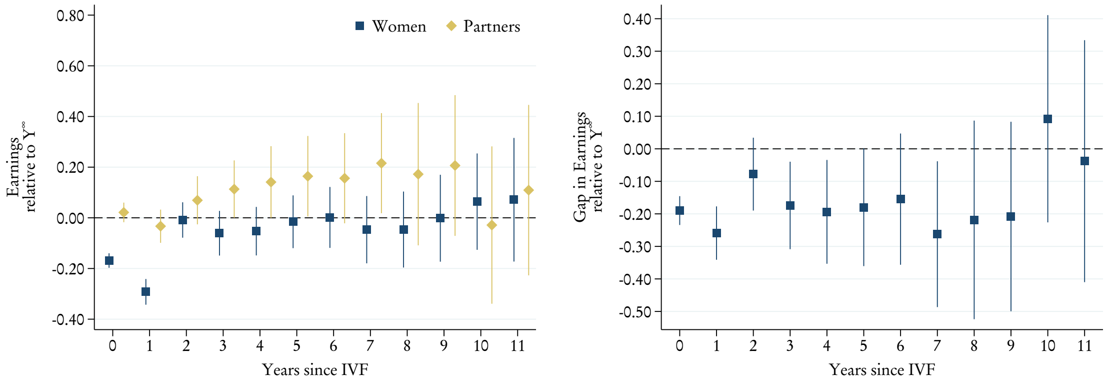
(event-IV). For each model, we report scaled estimates for women, partners, and the gap between the two (women minus partners, as in equation 12). Our discussion of long-term effects will focus on age 6, as our panel is balanced up to and including this age. However, we also present effects up to age 11, though most of these results are confined to the appendix. Results for other labor market outcomes are reported in the appendix.

5.1 Event-study estimates

We start by reporting the results using the regular event-study specification of equation (2), estimated on IVF-women and their partners in Figure 3(a).[16] Both women and partners display a comparable pre-trend leading up to birth, indicating that those who have children earlier are on relatively steeper age-earnings profiles compared to those who have children later. Following birth, IVF women see a sharp drop in earnings of about 27 percent which then attenuates somewhat and stabilizes at around 13 percent in the longer run. Partners, in contrast, experience almost no negative effects on earnings following childbirth. Rather, they see a small increase of about 2 percent in the longer-run.

Figure 3(b) shows the corresponding effects on the earnings gap between women and their partner. As both parents follow a similar upward-sloping trend in earnings there is

[16]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. For completeness, we also report results later in the paper using the event-study estimator proposed by Callaway and Sant'Anna (2021) (Figure A5), which relaxes the cohort-homogeneity assumption.



(a) Women and partners

(b) Gap

Figure 4. Earnings – LPR-IV estimates

Note: Estimated effects of fertility on earnings using the LPR-IV model described in equation (4) on our data. Panel (a) shows effects separately for women and partners, panel (b) shows the difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in Section 3.5. Full set of point estimates are reported in Table A8 in the appendix. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; women = 10,033). Results for other labor market outcomes are reported in appendix Figures A7 and A8.

no discernible pre-trend, but there is still a substantial difference after birth at around 15 percent, which in the long-run is almost entirely driven by the drop in women's earnings.

We repeat our analysis on a sample of non-IVF women in Section 7, and find similar results. The estimates for IVF women are therefore in line with existing event-study evidence from Norway (Andresen and Havnes, 2019; Andresen and Nix, 2022) and comparable countries such as Denmark (**?**).

5.2 LPR-IV estimates

We now turn to the estimated earnings effects of fertility using the LPR-IV model described in equation (4) with the outcome of the IVF treatment as the instrumental variable. Appendix Figure A6 reports the estimated first stages from equation (4), essentially the difference between the average fertility rates between successful and failed IVF attempts shown in Figure 1. By construction, the first stage equals 1 nine months after the IVF treatment. It then declines over time as always-takers realize fertility. In our annual data the first stage coefficient is already below 0.8 by the end of the first year, before stabilizing at 0.2 in the longer run. Despite this decline, the estimates are all highly statistically significant: Women who are successful in their first IVF-trial are therefore always more likely to have children than those who failed their first trial. The F-statistic is never below 500 in the first nine years since IVF and are reported in appendix Table A9.

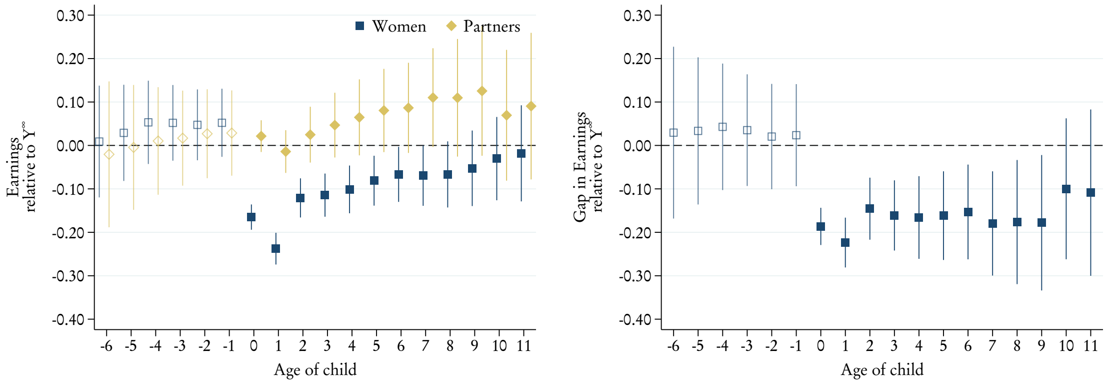
Figure 4(a) shows the IV estimates of equation (4), separately for women and their partners. Women's earnings drop by about 30 percent in the year following the IVF treatment, but the effect quickly reverts to zero in the second year, at which level it remains for the remaining period. For comparison, Lundborg et al. (2017) find long run earnings losses for women at around 11 percent. Our flatter profile may reflect differences in setting, sample period, or estimation horizon. As discussed in Section 3, this estimate is probably biased toward zero (i.e. the actual effect is more negative than the estimate) since delayed fertility is confounding the counterfactual earnings profile and introduces a positive bias.

Partners see no earnings drop immediately following IVF treatment. If anything, there is an earnings premium of about 16 percent six years after the IVF attempt. While the estimates are increasingly noisy they appear to be stable, estimating the average impact for a three year window around year six gives an estimate that is significant at conventional levels. Figure 4(b) reports the estimated effect of fertility on the earnings gap between women and their partners. This fluctuates a bit over time, averaging at 15 percent after six years, driven exclusively by the positive point estimate for partners' earnings.

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

5.3 Event-IV estimates

Figure 5 presents the estimated effects of children from our event-IV specification as described in equation (10, event-IV). F-statistics for the first stages are reported in appendix Table A9 and far exceed conventional levels for statistical significance. In Figure 5(a) we see that while we estimate a drop in women's earnings of about 24 percent when having a one-year-old, the longer run earnings loss is around 7 percent for a six-year-old. This is half of the effect on earnings estimated in the event-study model and the differences are statistically different at conventional significance levels. We also see no signs of any anticipation effects in the years leading up to the trial. No meaningful earnings drop is seen for partners around childbirth. By contrast, the estimates suggest an increase in earnings over time, reaching around 9 percent when the child is six. 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, after which it drops to around 22 percent, before stabilizing at around 15 percent. In the very long run the parental earnings gap is primarily driven by the



(a) Women and partners

(b) Gap

Figure 5. Earnings – Event-IV estimates

Note: Estimated effects of age of child on earnings using the event-IV model described in equation (10, event-IV). Panel (a) shows effects separately for women and partners, panel (b) shows the difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 3.5. Full set of estimates in table form are reported in Table A10 in the appendix. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; women = 10,033). Results for other labor market outcomes are reported in appendix Figures A9 and A10

partners.

For completeness, appendix Figures A9 and A10 report results for additional labor market outcomes (hours, employment and hourly wages) for the event-IV model.[17] The broad takeaway from the event-IV estimates is that for women the results on the long run seem to be mostly driven by responses at the employment margin. While disentangling intensive and extensive margin responses is not possible without a structural model, the results in Figure A10, which condition on employment, suggest that while women reduce hours on the short-run, their hours responses on the longer run appear to be negligible. Similar results for hourly earnings also give little sign that there are sizeable long-run effects. For partners, we also find employment responses. Contrary to women, the results in Figure A10 suggest that there are some long-run impacts on hours and hourly earnings, where the latter may be explained by career returns.

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

[17] Results for the same outcomes for the other models are reported in Figures A3 - A8.

estimates.

5.4 Economic interpretation of partner and household responses

Children reduce maternal earnings but raise partner earnings substantially, so the household-level incidence of children differs from what maternal earnings alone would suggest. A natural interpretation is that children increase both childcare demands and household income needs, inducing a reallocation within the couple in which mothers shift more time into childcare while partners shift further into market work. Partner earnings can then rise either because partners work more to finance the additional costs of children, or because specialization lets them invest more continuously in market work and move onto a steeper earnings path (Becker, 1981, 1985). This interpretation may be reinforced by bargaining frictions and gender norms (Chiappori, 1992; Lundberg and Pollak, 1996). An additional possibility is that fathers receive a labor-market premium independently of their own behavioral response, because employers perceive them as more committed or stable workers (Lundberg and Rose, 2000, 2002). Our design does not separately identify these channels.

Finally, our estimates are post-birth by design and do not capture earnings adjustments during the broader fertility journey before the IVF success shock is realized. If partners reduce labor supply during that period and increase it after birth or after fertility uncertainty is resolved, part of the estimated increase in partner earnings may reflect that broader transition into parenthood rather than the post-birth allocation of time alone. This may matter in particular for IVF couples, for whom the timing and uncertainty of conception differ from natural conceptions.

6 Reconciling estimates of the effect of fertility

Table 3 summarizes the estimates for the three models by reporting the long-run estimates of earnings for the woman, the partner, and the gap between the two known as the child penalty. Long–run estimates are evaluated when the child is six years old ($a = 6$) which is the last age for which we have a balanced panel. We report analogous results for age eleven ($a = 11$) in appendix Table A11.[18] Column (1) shows estimates from the LPR-IV model, column (2) shows estimates from the event model, column (3) shows estimates from the event-IV model. The final column compares the estimates from the event model with the event-IV model. This difference is informative about timing-related bias in the event estimates under the assumptions of the event-IV model and absent notable complier heterogeneity, which we discuss below.

[18] Appendix Tables A12 and A13 show that removing the restriction to partnered women in the year before IVF gives estimates nearly identical to the baseline.

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

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

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

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

The estimate for the long-run parental earnings gap from the event-IV specification is identical to that of the LPR-IV model. However, when it comes to the separate estimates for women and partners, the event-IV model paints a different picture than the event-study model. For women, it estimates only a small long-run negative impact of children on earnings of 7 percent, compared to 13 percent in the event-study model. For partners, the event-IV model estimates an earnings increase of around 9 percent compared to 2 percent in the event-study model.

The estimates for age 11 in Table A11 paint a very similar picture. The estimated gaps in the event models appear to be stable, and the event-IV estimates suggest that the effect for women is smaller, while the partner effects remain positive or are even increased. Note however that our sample is no longer balanced and is much reduced. The estimates are consequently very noisy, which means that we cannot reject that the effects are the same as for age 6.

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

6.1 Event-IV and LPR-IV

We report the estimated weights in (9) for $p = 6$ in Figure 6. On the left-hand y-axis we plot the first-stage coefficients for having a child of age a six years after IVF (π_{a6}). The right-hand y-axis shows the normalized weight for each first-stage (ω_{a6}). The figure shows that there is a large positive weight for $a = 6$ which means that when estimating the fertility effect on earnings, the LPR-IV estimator puts a large positive weight on the effect of having a child p years old. However, the estimated effects for having a child any younger than six years old (i.e. $a < p$) are given a negative weight. When effects are most negative for younger children, this weighting pushes the fertility estimates in the LPR-IV model towards zero relative to the contemporaneous effect of having a child within a year from the IVF attempt, which has a positive weight. We show that this pattern holds for all p in appendix Figure A12 and A13. On the very short run ($p = 0$) the fertility effect γ_0 is equal to the earnings effect δ_0, but over time γ_p reflects not only the contemporaneous earnings effect δ_p, but also increasingly more mass on younger-child effects that enter with negative weights.

We can use our event-IV estimates of δ_a and π_{ap} to construct alternative estimates of γ_p and compare these to the estimates of γ_p based on the LPR-IV estimates from equations (4) and (5). The mapping is illustrated in appendix Figure A14 where we plot the results for women's earnings from the LPR-IV model along with the estimates constructed from the reduced form and the first stages from our event-IV. These results are consistent with the equivalence between the reduced forms and show that the difference comes from decomposing fertility into dynamic treatment effects by child age.

6.2 Event-IV and Event

If counterfactual earnings trajectories are correlated with when women choose to start trying for a child, then this could violate the event-study assumption (3), which states that fertility timing should be mean independent of counterfactual earnings conditional on age and time. One way to investigate this is by adjusting for fertility intentions. Our IVF data provide unusually direct information about intended fertility timing, since we observe the date of the first IVF attempt and can therefore compare women by when they were trying to conceive.

This lets us assess how sensitive standard event-study estimates are to conditioning on

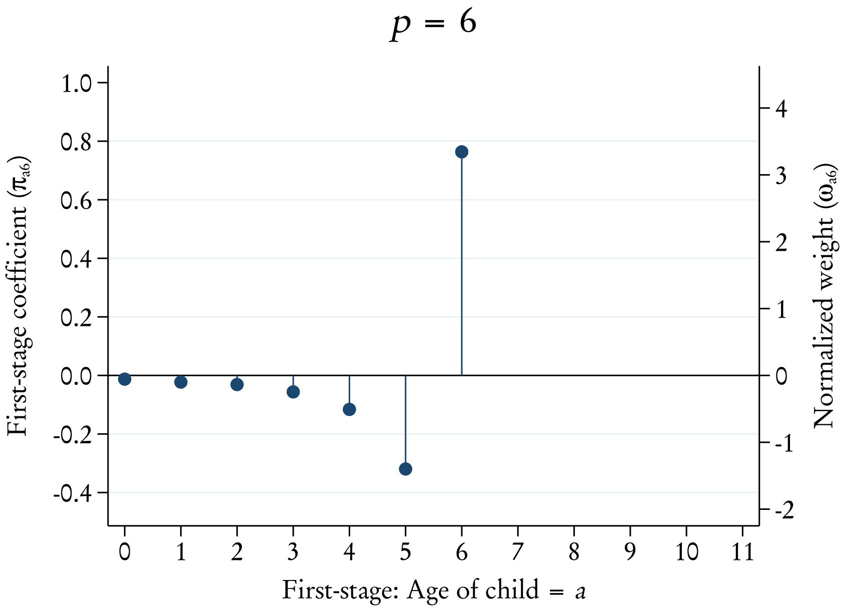


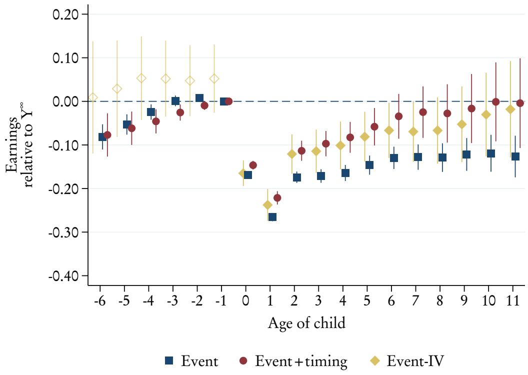
Figure 6. Mapping the first stages of LPR-IV to event-IV

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

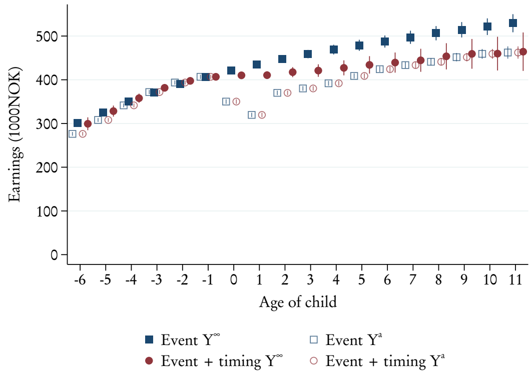
observed fertility-attempt timing, where the attempt date is used as conditioning information rather than as a source of exogenous variation. Controlling for timing within an event-study framework – still within the IVF sample – helps separate the role of timing from other differences between the event-study and IV approaches, such as the latter being restricted to compliers of the first IVF treatment.

Figure 7a compares the standard event study to the same specification augmented with dummies for time since the first IVF trial. Controlling for timing has little impact on pre-trends, but substantially reduces the estimated post-birth effects on women's earnings. Where the earnings reduction for women was about 13 percent in the standard event-study setup, controlling for timing lowers the estimated penalty to about 3 percent after 6 years, and close to zero by year 11.

Figure 7b provides further insight by reporting estimated counterfactual earnings, normalized at $a = -1$, for the event-study with and without controlling for timing. The pre-birth earnings profiles are very similar across the two specifications, consistent with the small effect of timing controls on pre-trends in Figure 7a. The main difference arises after birth: once we control for fertility-attempt timing, the estimated counterfactual earnings



(a) Event vs. Event-IV estimates



(b) Counterfactual earnings profiles – Event-study estimates

Figure 7. Event-study and Event-IV estimates

Note: Panel (a) 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 in section 3.5. Panel (b) shows estimated potential earnings without (Y^{∞}) and with child (Y^{a}) from the event-study model, with and without IVF-timing controls. The IV sample includes women with a first IVF treatment between 2009–2016, no prior children, and a registered partner (obs. = 173,480; women = 10,033); the event-study sample restricts to those who eventually had a child (obs. = 145,571; women = 8,349).

profile Y^{∞} is flatter. In the IVF sample, first births therefore occur when counterfactual earnings growth begins to slow, a pattern that is not captured in the standard event-study specification without timing controls.

We next compare the event-study estimates with timing controls to the full event-IV estimates. Figure 7a shows that once we control for timing in the event-study model, the estimated fertility effects move much closer to the event-IV estimates, and the remaining differences are no longer statistically significant. Observed fertility-attempt timing accounts for an important share of the gap between the standard event-study and event-IV estimates in the IVF setting. The remaining differences reflect that the two approaches rely on different identifying variation and different local populations.

The appendix reports two supplementary event-study exercises: the treatment-cohort estimator of Callaway and Sant'Anna (2021) (Figure A5) and the pre-trend extrapolation exercise of Rambachan and Roth (2023) (Figure A17). Both point in the same direction as the main timing result. In the IVF sample, the alternative estimators line up the pre-trends but imply larger negative earnings effects than either the event-study specification that conditions on observed fertility-attempt timing or the event-IV estimates. This reinforces the narrower point that, in this setting, the presence or absence of pre-trends can be misleading about the sign of post-birth selection bias.

Taken together, these results are consistent with endogenous fertility timing generating upward bias in estimated maternal earnings losses and, correspondingly, in the maternal contribution to the earnings gap. More specifically, within the IVF sample they suggest that the exogeneity assumption underlying the standard event-study specification conditional on age and time fails. However, these findings should be interpreted within the IVF context. The comparison is between women who are actively trying to conceive and whose fertility-attempt timing we observe directly. It is therefore not clear that the findings would translate to the general population in Norway, where many pregnancies are also planned, but where fertility-attempt timing is typically unobserved and the process of conception is often less prolonged, medically intensive, and uncertain than in IVF.

7 External validity

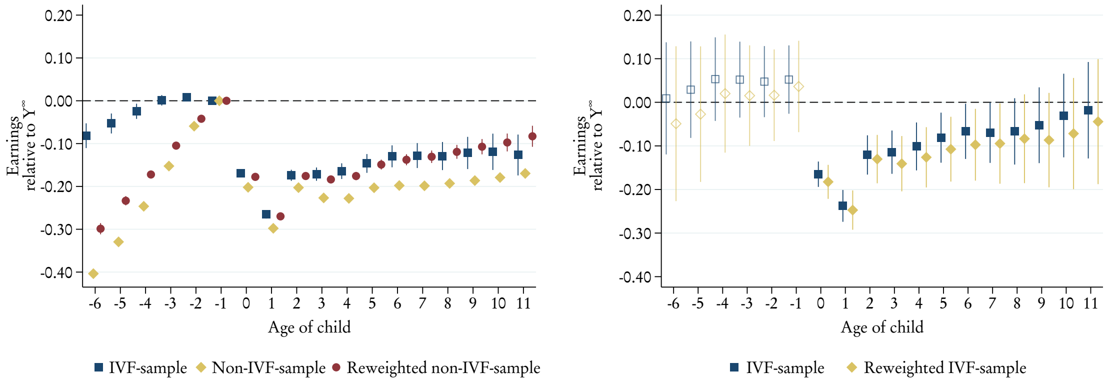
Our results are based on IVF women, who make up around six percent of all births in Norway. This is a sizable and growing group, and therefore of interest in its own right. The advantage of this setting is that it allows us to exploit two key features uniquely provided by the IVF context: i) detailed information about the timing of fertility and ii) the maintained assumption that IVF success is random conditional on women's age and education. However, broader implications outside the IVF setting are necessarily more tentative. Generalization rests on two questions: (i) whether IVF women differ from other women, and (ii) whether IVF births differ from other births. We discuss these in turn.

7.1 Are IVF women different from non-IVF women?

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

We start by comparing fertility timing across samples. Table 1 shows that women undergoing IVF are older and more educated at conception and, on average, have slightly fewer children. These intensive-margin differences are fully accounted for by age and education. When we reweight the non-IVF sample to match the IVF distribution of age and education at conception, completed fertility among those with at least one child is virtually identical across groups (Table A3).[19] Table A4 reports additional characteristics for the reweighted non-IVF sample. Conditional on age and education, IVF women (and their partners) have somewhat higher earnings, and IVF women exhibit slightly higher sickness

[19]Observations are weighted by the inverse propensity score, with scores estimated via a probit fully saturated in age and education indicators. Women outside the common support of the IVF sample on age and education (n=1,005) are excluded from the reweighted sample.



(a) Event-study specification

(b) Event-IV specification

Figure 8. Earnings. Reweighted estimates

Note: Panel (a) shows event-study estimates for the IVF sample (145,571 observations; 8,349 women), the non-IVF sample (1,972,754 observations; 109,791 women), and the non-IVF sample reweighted to match the composition of the IVF group (108,786 women). All samples include women who eventually have at least one child. Panel (b) shows event-IV estimates for the IVF sample and for the same sample reweighted to match the composition of non-IVF women. The event-IV sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children beforehand, and were registered with a partner at the time (173,480 observations; 10,033 women). Weights are inverse propensity scores estimated from a fully saturated probit model in age and education at first birth. The reweighted non-IVF sample excludes 1,005 women who fall outside the common support of the IVF sample on these characteristics. Estimates are scaled relative to each gender's counterfactual outcomes (Y^{∞}), as described in section 3.5.

absence. IVF couples therefore appear modestly positively selected, but the remaining differences are small.

An alternative way to assess population heterogeneity is to compare standard event-study estimates for IVF women with those for non-IVF women, and to diagnose differential selection or anticipation from pre-trends. Figure 8a reports estimates for both groups. Pre-trends are noticeably steeper in the non-IVF population, consistent with stronger timing selection. This is in line with evidence from the Norwegian Mother and Child Cohort Study (Magnus et al., 2006), where 82 percent of mothers report a planned pregnancy, implying substantial scope for unconstrained positive timing of births in the general population. At the same time, the remaining unplanned pregnancies may introduce additional heterogeneity and negative selection. This is consistent with Gallen et al. (2023), who, using Swedish data, find larger negative effects for younger women. In contrast, IVF conceptions are actively planned and timing is more constrained by treatment schedules, leaving less scope for selective timing.

A comparison of the post-birth fertility effects can also be informative about heterogeneity across populations. Figure 8a shows that non-IVF women experience a larger drop in earnings following birth than IVF women. However, reweighting the non-IVF sample makes the event-study estimates of responses to children remarkably similar to those of the

IVF sample. This suggests that, once the non-IVF sample is aligned on age and education, observed post-birth earnings dynamics are similar across the two groups. Since event-study estimates remain affected by endogenous timing, this comparison is descriptive and does not by itself establish that the underlying causal responses are the same.

To examine whether causal effects are heterogeneous by age and education, we additionally reweight the IVF sample to match the composition of non-IVF women before re-estimating the event-IV model. Since selection bias is eliminated in these specifications, any observed differences should reflect treatment effect heterogeneity alone. Figure 8b shows that this adjustment only marginally changes the estimated impacts of children. While impacts are somewhat larger for less educated and younger women, differences in age and education between the IVF sample and the overall population appear to have little effect on the causal estimands we report.

Reweighting on age and education makes the event-study profiles in the IVF and non-IVF samples more similar, and reweighting the IVF sample to match non-IVF women changes the event-IV estimates only slightly. These exercises show that observable differences in age and education do not account for much of the gap between the IVF event-IV estimates and the non-IVF event-study estimates.

Beyond observable characteristics, there could be other differences that may complicate direct comparisons. Women seeking fertility treatment may make labor market adjustments in the years leading up to conception that women in the general event-study sample pursue only to a lesser degree, and such pre-IVF adjustments would not be captured by the IVF design. It is also important to emphasize that because the underlying variation in fertility differs, there may be differences in partners' earnings profiles as well. Partners may, for instance, postpone promotions in anticipation of having children to a larger extent than partners of women in the event-study sample, where conception is on average achieved faster than among IVF women.

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

A frequently noted phenomenon in the context of IVF is the presence of a "disappointment effect" following an unsuccessful IVF, which could affect behavior or mental health (e.g. Gallen et al., 2023; Bögl et al., 2024; Martinenghi and Naghsh-Nejad, 2025). In our design, the effects of disappointment from not conceiving and the effects of having a child on mental health and marital stability cannot be separately identified, as they are opposite realizations of the same IVF attempt.

In our data, mental-health outcomes differ around treatment. In particular, in the short run the probability of having at least one mental-health–related medical visit per year is about 12 percentage points (baseline mean 14) lower for those who become mothers than for those who do not (Figure A15a). The probability of having at least one *severe* psychological diagnosis declines by about 3 percentage points (baseline mean 6) among mothers relative to non-mothers, an effect that is statistically significant only in the first year after childbirth. In the longer run, we do not detect effects on these mental-health outcomes. The sensitivity checks in Figure A16 show that the estimated earnings effects change little when we exclude potentially affected women.

Impacts on other non–labor-market outcomes, such as marital stability, are modest in the data. Figure A15b shows that having children reduces divorce risk over time by 5 percentage points relatively to the baseline where everyone was partnered.

Finally, the age-of-child estimates are nearly invariant as the estimation horizon expands (Figure A2). Extending the horizon necessarily brings in comparisons involving childless women observed further from an unsuccessful first IVF attempt, and often after additional unsuccessful attempts, so any large and persistent disappointment-related effects among unsuccessful women would be expected to generate systematic drift in the estimated age profile as the horizon increases. We do not observe such drift.

These patterns do not resolve whether disappointment-related channels matter for identification. In our design, disappointment from not conceiving and the effects of having a child are two sides of the same coin, so the two cannot be separately identified. Mental-health and relationship responses may therefore proxy for broader unobserved changes following IVF failure, and we treat this as a limitation of the design.

8 Conclusion

Social scientists and policy makers have devoted considerable effort to understanding the drivers of the gender wage gap. Much of this work studies how parenthood, and especially motherhood, contributes to that disparity. A broad conclusion is that women's earnings fall abruptly and persistently after first birth, whereas their partners' earnings remain largely unchanged. The resulting increase in the earnings gap between mothers and fathers is commonly referred to as the child penalty. Empirically, much of this literature relies on event-study designs.

This paper compares three estimators within the IVF setting: the standard event study, the IVF-based IV estimator of Lundborg et al. (2017), and an event-IV estimator that uses IVF success to recover age-of-child effects. These approaches rely on different identifying variation and target different estimands. The standard event study is birth-centered and targets age-of-child effects under exogenous timing assumptions. LPR-IV is IVF-attempt-centered and identifies fertility effects conditional on trying to conceive. Event-IV places the IV and event-study approaches on a common child-age scale within the IVF sample.

Using data on Norwegian women undergoing IVF treatment, we find that first births occur when estimated counterfactual no-child earnings growth starts to flatten. Within the IVF sample, conditioning on the timing of the first fertility attempt therefore substantially reduces the maternal earnings loss implied by the standard event study. Using IVF success to instrument for fertility addresses remaining endogeneity within the IVF setting, but does not overturn that conclusion. Our preferred event-IV estimates imply a longer-run earnings loss for mothers of about 7 percent, roughly half the effect implied by the standard event-study specification.

We also clarify how the IVF-based estimator of Lundborg et al. (2017) relates to birth-centered event studies. The LPR-IV estimator mixes effects of children of different ages and, over time, places increasing negative weight on births occurring after the first IVF attempt. Event-IV instead recovers age-of-child effects within the IVF design, making the IV and event-study approaches directly comparable within the IVF sample.

The overall earnings gap between women and partners is similar across the three models we study, but the implied incidence differs. The LPR-IV estimates attribute the gap to higher partner earnings, while the standard event-study estimates attribute most of the gap to lower maternal earnings. In the event-IV estimates, by contrast, the gap reflects both lower earnings for women and higher earnings for partners. Similar child penalties can therefore mask different incidence across mothers and partners, and the partner response matters for how the child penalty should be interpreted.

The paper documents a form of selection that is natural in economic models of fertility timing and that can materially affect estimates when fertility-attempt timing is unobserved. We establish the magnitude of this mechanism within the IVF setting. Whether a similar discrepancy between standard event-study and event-IV estimates extends to non-IVF populations remains an open question, since those populations can differ in fertility intent, anticipatory labor market behavior, and exposure to disappointment effects.

References

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

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

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

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

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

Angrist, J. D., B. Ferman, C. Gao, P. Hull, O. L. Tecchio, and R. W. Yeh (2025). Instrumental variables with time-varying exposure: New estimates of revascularization effects on quality of life. Working paper no. 33296, NBER.

Becker, G. S. (1981). *A Treatise on the Family*. Cambridge, MA: Harvard University Press.

Becker, G. S. (1985). Human capital, effort, and the sexual division of labor. *Journal of Labor Economics 3*(1), S33–S58.

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

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

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

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

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

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

Chiappori, P.-A. (1992). Collective labor supply and welfare. *Journal of Political Economy 100*(3), 437–467.

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

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

Gallen, Y., J. S. Joensen, E. R. Johansen, and G. F. Veramendi (2023). The labor market returns to delaying pregnancy. Technical report. Revise and Resubmit, *American Economic Review*.

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

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

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

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

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

Ketel, N., E. Leuven, H. Oosterbeek, and B. van der Klaauw (2012). The returns to medical school in a regulated labor market: Evidence from admission lotteries. Working paper, Universiteit van Amsterdam. September 2012.

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

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

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

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

Kunze, A. and M. Ejrnæs (2004). Wage dips and drops around first birth. Technical Report IZA DP No. 1011, Institute for the Study of Labor (IZA), Bonn.

Lundberg, S. and R. A. Pollak (1996). Bargaining and distribution in marriage. *Journal of Economic Perspectives 10*(4), 139–158.

Lundberg, S. and E. Rose (2000). Parenthood and the earnings of married men and women. *Labour Economics 7*(6), 689–710.

Lundberg, S. and E. Rose (2002). The effects of sons and daughters on men's labor supply and wages. *Review of Economics and Statistics 84*(2), 251–268.

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

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

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

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

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

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

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

A Appendix For Online Publication

A.1 Mapping LPR-IV to child-age

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

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

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

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

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

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

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

The paper presents diagnostic evidence consistent with this assumption.

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

Monotonicity,

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

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

By definition,

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

and the age-horizon specific first-stage equals:

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

since

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

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

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

Fertility p years since IVF is defined as

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

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

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

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

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

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

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

Therefore the reduced form

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

equals

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

This shows that

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

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

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

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

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

This decomposition also clarifies the horizon-specific objects entering the reduced form. At each horizon p, the age-a first stage is generated by the same subpopulation that defines the corresponding treated and untreated levels in δ_{ap}. For $a < p$, this subpopulation consists of women for whom a failed first IVF attempt would have delayed birth until period $p-a$, that is, women with $P_i(0) = p-a$. For $a = p$, the relevant comparison is the full sample, since success mechanically implies a birth in period 0. Thus

$$\delta_{ap} = \mu_{ap}^a - \mu_{ap}^\infty,$$

where

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

Without further restrictions, the equation at horizon p provides one reduced-form relation in the $p + 1$ unknown objects $\delta_{0p}, \ldots, \delta_{pp}$, so the horizon-specific effects δ_{ap} are not all identified. Recovering a single age-specific effect δ_a in the recursive/event-IV system therefore requires the timing-homogeneity restriction

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

If one also wants a single age-specific untreated level in the stacked system, an additional counterfactual-level homogeneity restriction is needed:

$$\mu_{ap}^{\infty} = \mu_a^{\infty} \qquad \text{for all } p \geq a.$$

Together with timing-homogeneity, this implies treated-level homogeneity as well:

$$\mu_{ap}^{a} = \mu_{ap}^{\infty} + \delta_{ap} = \mu_a^{\infty} + \delta_a \equiv \mu_a^{a}.$$

The next corollary shows how the treated levels can be identified horizon by horizon using suitable outcome transformations, and how the untreated levels can be recovered once the relevant age effects have been identified through the recursive system.

Corollary: identifying treated and untreated levels for the same subpopulations.

Define the transformed outcomes

$$\tilde{y}_{ip}^{\infty,a} \equiv -y_{ip}\mathbf{1}\{a_{ip} \neq a\}, \qquad \tilde{y}_{ip}^{a} \equiv y_{ip}\mathbf{1}\{a_{ip} = a\}.$$

Using

$$y_{ip} = y_{ip}^{\infty}\mathbf{1}\{a_{ip} = \infty\} + \sum_{b=0}^{p} y_{ip}^{b}\mathbf{1}\{a_{ip} = b\},$$

we can rewrite

$$\tilde{y}_{ip}^{\infty,a} = -y_{ip}^{\infty} + y_{ip}^{\infty}\mathbf{1}\{a_{ip} = a\} - \sum_{b \neq a}(y_{ip}^{b} - y_{ip}^{\infty})\mathbf{1}\{a_{ip} = b\}.$$

Applying the same steps as in the derivation of ρ_p, the reduced form for $\tilde{y}_{ip}^{\infty,a}$,

$$\tilde{\rho}_p^{\infty,a} \equiv E[\tilde{y}_{ip}^{\infty,a} \mid \text{success}_i = 1] - E[\tilde{y}_{ip}^{\infty,a} \mid \text{success}_i = 0],$$

equals

$$\tilde{\rho}_p^{\infty,a} = \mu_{ap}^{\infty}\pi_{ap} - \sum_{b \neq a} \delta_{bp}\pi_{bp}.$$

Hence, provided $\pi_{ap} \neq 0$,

$$\mu_{ap}^{\infty} = \frac{\tilde{\rho}_p^{\infty,a} + \sum_{b \neq a} \delta_{bp}\pi_{bp}}{\pi_{ap}}.$$

For the treated level, note that

$$\tilde{y}_{ip}^{a} = y_{ip}^{a}\mathbf{1}\{a_{ip} = a\}.$$

The corresponding reduced form,

$$\tilde{\rho}_p^{a} \equiv E[\tilde{y}_{ip}^{a} \mid \text{success}_i = 1] - E[\tilde{y}_{ip}^{a} \mid \text{success}_i = 0],$$

satisfies

$$\tilde{\rho}_p^{a} = \mu_{ap}^{a}\pi_{ap},$$

so that, provided $\pi_{ap} \neq 0$,

$$\mu_{ap}^{a} = \frac{\tilde{\rho}_p^{a}}{\pi_{ap}}.$$

Thus the treated levels μ_{ap}^{a} are directly identified horizon by horizon, while the untreated levels μ_{ap}^{∞} can be recovered horizon by horizon once the relevant age effects have been recovered from the recursive system. Without further restrictions, the horizon-specific objects δ_{ap} are not all identified. Under timing-homogeneity, however, the recursive argument identifies the common age-specific effects δ_a, which in turn allow recovery of the corresponding untreated and treated levels for the same horizon-specific subpopulations. In the empirical implementation, we estimate the corresponding transformed outcomes in the stacked event-IV system, but the identification result itself is the horizon-specific recursive result shown above.

This transformed-outcome approach builds on Abadie (2003), who shows how related transformations recover complier means under a binary instrument. Here we derive the corresponding result for the horizon-specific age-by-horizon cells of our event-IV setup.

Estimation Consider the transformed-outcome event-IV specification we estimate

$$\tilde{y}_{it}^{\infty,a} = \sum_{b\geq 0} \beta_b^{\infty,a} \mathbf{1}\{a_{it} = b\} + x_{it}'\psi + \tau_t + \sum_{p\geq 0} P_{it,p}\theta_p + \varepsilon_{it}^{\infty,a},$$

instrumenting each age indicator $\mathbf{1}\{a_{it} = b\}$ with the same event-IV instruments success$_i P_{it,p}$ used in the main specification. As in the main text, stacking horizons in calendar time implies that the reduced-form coefficient on the instrument success$_i P_{it,p}$ equals the corresponding horizon-p reduced form.

For the untreated transformation, the horizon-specific reduced form is

$$\tilde{\rho}_p^{\infty,a} = \mu_a^{\infty} \pi_{ap} - \sum_{b\neq a} \delta_b \pi_{bp}.$$

In the stacked transformed-outcome regression, the coefficient on the age-b indicator contributes π_{bp} to the reduced form at horizon p. Hence the reduced-form coefficient on success$_i P_{it,p}$ implied by the stacked specification is

$$\sum_{b=0}^{p} \beta_b^{\infty,a} \pi_{bp}.$$

Matching this to the horizon-specific reduced form for every p gives

$$\beta_a^{\infty,a} = \mu_a^{\infty}, \qquad \beta_b^{\infty,a} = -\delta_b \quad \text{for } b \neq a.$$

Equivalently, because $\pi_{ap} = 0$ for $p < a$ and $\pi_{aa} = 1$, the equation at $p = a$ yields

$$\mu_a^{\infty} = \tilde{\rho}_a^{\infty,a} + \sum_{b=0}^{a-1} \delta_b \pi_{ba}.$$

Similarly, for the treated transformation, consider

$$\tilde{y}_{it}^{a} = \sum_{b\geq 0} \beta_b^{a} \mathbf{1}\{a_{it} = b\} + x_{it}'\psi + \tau_t + \sum_{p\geq 0} P_{it,p}\theta_p + \varepsilon_{it}^{a}.$$

The horizon-specific reduced form is

$$\tilde{\rho}_p^{a} = \mu_a^{a} \pi_{ap},$$

so matching reduced forms gives

$$\beta_a^{a} = \mu_a^{a}, \qquad \beta_b^{a} = 0 \quad \text{for } b \neq a.$$

Again, since $\pi_{ap} = 0$ for $p < a$ and $\pi_{aa} = 1$,

$$\mu_a^a = \tilde{\rho}_a^a.$$

Thus, under timing-homogeneity and counterfactual-level homogeneity, the stacked transformed-outcome event-IV system recovers both the untreated and treated levels,

$$\mu_a^\infty \qquad \text{and} \qquad \mu_a^a,$$

and therefore

$$\delta_a = \mu_a^a - \mu_a^\infty.$$

A.2 Standard errors on rescaled estimates

Denote the rescaled estimate by x:

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

The Delta method gives

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

where

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

since

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

we also have that

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

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

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

where $V(\delta), V(y^1)$ and $V(y^{\infty})$, all come from separate 2SLS regressions as outlined in section 3.4.

A.3 Additional results

Table A1. ICPC-2 Chapter P: Psychological Codes and Descriptions

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

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

Table A2. Descriptive statistics for IVF women and non-IVF women: full table

	IVF (1)	Non-IVF (2)	Difference (3)	
Woman characteristics				
Number of IVF attempts	2.84			
Success, first trial	0.31			
Success, endpoint	0.63			
Fertility, endpoint	0.83	1.00	-0.17	(0.00)
Total number of children	1.47	1.97	-0.50	(0.01)
0 children	0.17	0.00		
1 child	0.30	0.23	0.07	(0.00)
2 children	0.44	0.60	-0.15	(0.01)
3 children	0.09	0.16	-0.07	(0.00)
4 children	0.01	0.02	-0.01	(0.00)
Age	31.82	28.40	3.41	(0.05)
Education				
- Compulsory	0.14	0.17	-0.03	(0.00)
- High School	0.24	0.23	0.01	(0.00)
- Bachelor	0.42	0.41	0.01	(0.01)
- Master	0.20	0.19	0.01	(0.00)
Earnings (1000 NOK)	362.72	285.54	77.18	(1.86)
Hours (FTE)	0.88	0.78	0.10	(0.00)
Employed	0.80	0.67	0.14	(0.00)
Hourly earnings (NOK)	221.12	197.45	23.68	(1.85)
Sickness absence days	14.98	10.13	4.85	(0.30)
Visits to general practitioner (GP)	2.51	2.01	0.50	(0.02)
Any (mild/severe) psychological symptoms	0.14	0.12	0.02	(0.00)
Severe psychological symptoms	0.06	0.06	0.00	(0.00)
Hospital days	2.13	0.88	1.25	(0.04)
Partner characteristics				
Age	35.07	31.21	3.86	(0.06)
Female	0.01	0.01	0.01	(0.00)
Education				
- Compulsory	0.17	0.19	-0.02	(0.00)
- High School	0.39	0.37	0.01	(0.01)
- Bachelor	0.27	0.26	0.01	(0.00)
- Master	0.17	0.17	0.00	(0.00)
Earnings (1000 NOK)	457.05	384.39	72.65	(2.90)
Hours (FTE)	0.85	0.78	0.07	(0.00)
Employed	0.84	0.76	0.08	(0.00)
Hourly earnings (NOK)	281.24	254.61	26.64	(2.42)
N Women	10 033	109 791		

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

Table A3. Total number of children.

Number of children	Non-IVF sample (1)	Reweighted non-IVF sample (2)	IVF sample (3)
1	0.23	0.34	0.36
2	0.60	0.55	0.53
3	0.16	0.10	0.11
≥ 4	0.02	0.01	0.01
Number of women	109,791	108,786	8,346

Note: This table shows the number of children by the end of the sample period, conditional on having at least one child. Column 1 shows fertility for the non-IVF sample; column 2 for the non-IVF sample reweighted to match the distribution of the IVF sample; and column 3 for the subsample of the IVF sample which includes women with at least one child. The reweighted sample (column 2) excludes 1,005 non-IVF women who fall outside the common support of the IVF sample on age and education at first birth, and therefore receive no weight.

Table A4. Descriptive statistics for IVF women vs reweighted non-IVF women

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

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

Table A5. Complier characteristics

	All IVF women			Compliers	
	Mean (1)	Std.Dev. (2)		Mean (3)	Std.Dev. (4)
Woman characteristics					
Age	33.99	(4.21)		34.11	(4.25)
Pre-IVF earnings	27.05	(17.00)		26.77	(16.97)
Education					
- Compulsory	0.14	(0.35)		0.15	(0.36)
- High School	0.24	(0.43)		0.25	(0.43)
- Bachelor	0.42	(0.49)		0.41	(0.49)
- Master	0.20	(0.40)		0.19	(0.39)
Sickness absence days	32.23	(72.68)	36.84	(76.27)	
GP visits	3.51	(3.87)	3.81	(4.02)	
Any (mild/severe) psychological symptoms	0.15	(0.36)	0.16	(0.37)	
Hospital days	6.80	(29.78)	8.11	(34.09)	
Partner characteristics					
Age	35.06	(6.10)		35.33	(6.23)
Education					
- Compulsory	0.17	(0.37)		0.17	(0.38)
- High School	0.39	(0.49)		0.39	(0.49)
- Bachelor	0.27	(0.45)		0.27	(0.44)
- Master	0.17	(0.38)		0.16	(0.37)
Earnings	36.96	(24.36)		36.58	(23.65)
(Estimated) number of women	10,033			7,527.08	

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

Table A6. Descriptive statistics for IVF women by success at first trial

	Failure (1)	Success (2)	Difference (3)	
Woman characteristics				
Number of IVF attempts	3.31	1.81	-1.49	(0.03)
Success, endpoint	0.46	1.00	0.54	(0.01)
Total number of children	1.31	1.84	0.54	(0.02)
0 children	0.24	0.00	-0.24	(0.00)
1 children	0.29	0.30	0.00	(0.01)
2 children	0.38	0.58	0.20	(0.01)
3 children	0.08	0.12	0.04	(0.01)
4 children	0.01	0.01	0.00	(0.00)
Age	32.06	31.27	-0.79	(0.09)
Education				
- Compulsory	0.15	0.12	-0.03	(0.01)
- High School	0.24	0.23	-0.01	(0.01)
- Bachelor	0.41	0.44	0.03	(0.01)
- Master	0.20	0.21	0.01	(0.01)
Earnings (1000 NOK)	362.76	362.63	-0.13	(4.14)
Hours (FTE)	0.88	0.88	0.00	(0.01)
Employed	0.80	0.81	0.01	(0.01)
Hourly wage (NOK)	221.46	220.39	-1.07	(4.44)
Sickness absence days	15.11	14.69	-0.42	(0.73)
Visits to general practitioner	2.53	2.47	-0.06	(0.05)
Any (mild/severe) psychological symptoms	0.14	0.14	-0.00	(0.01)
Severe psychological diagnosis	0.06	0.07	0.00	(0.00)
Hospital days	2.21	1.95	-0.27	(0.15)
Partner characteristics				
Age	35.33	34.50	-0.83	(0.13)
Female	0.01	0.02	0.00	(0.00)
Education				
- Compulsory	0.17	0.16	-0.01	(0.01)
- High School	0.39	0.37	-0.02	(0.01)
- Bachelor	0.27	0.29	0.03	(0.01)
- Master	0.17	0.18	0.01	(0.01)
Earnings (1000 NOK)	457.17	456.79	-0.38	(6.34)
Hours (FTE)	0.85	0.85	0.00	(0.01)
Employed	0.83	0.84	0.01	(0.01)
Hourly wage (NOK)	280.70	282.41	1.71	(5.12)
N Women	6 881	3 152		

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

Table A7. Point estimates from the event study model.

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

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

Table A8. Point estimates from the LPR-IV model.

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

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

Table A9. First-stage F-statistics for LPR-IV and event-IV.

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

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

Table A10. Point estimates from the Event-IV model.

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

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

Table A11. Comparison of long-run (age 11) fertility effects and child penalty estimates across models

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

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

Table A12. Robustness: Comparison of long-run (age 6) fertility effects and child penalty estimates across models, without restriction that woman has partner

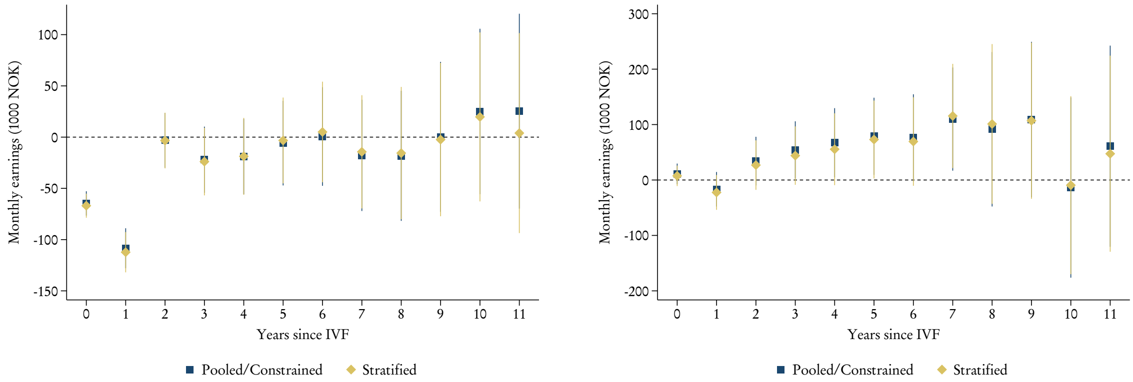
	LPR-IV (1)	Event (2)	Event-IV (3)	Event vs. Event-IV (2) - (3)
Gap	-0.07	-0.15	-0.11	-0.04
	(0.10)	(0.02)	(0.06)	(0.06)
Women	0.04	-0.14	-0.05	-0.09
	(0.06)	(0.01)	(0.03)	(0.03)
Partners	0.11	0.01	0.06	-0.05
	(0.10)	(0.02)	(0.06)	(0.06)

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

Table A13. Robustness: Comparison of long-run (age 11) fertility effects and child penalty estimates across models, without sample restriction that woman has partner

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

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

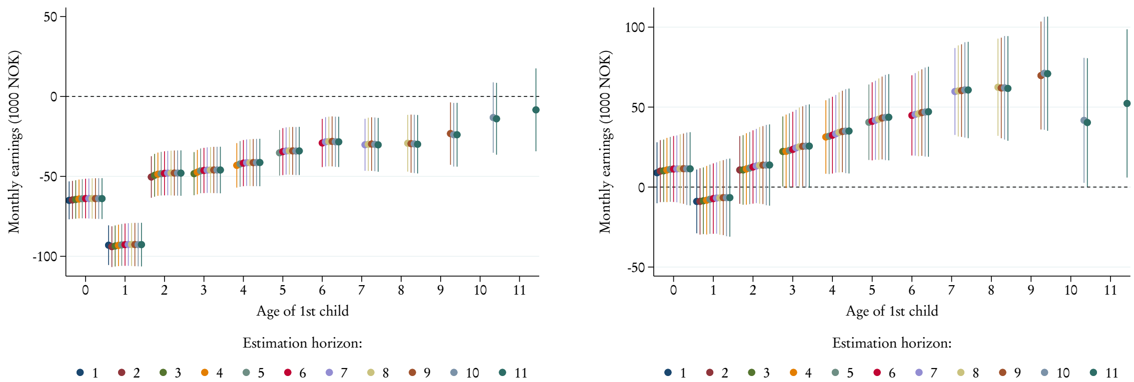


(a) Woman's earnings

(b) Partner's earnings

Figure A1. Results by potential age of child

Note: Estimated effects of fertility on earnings using the LPR-IV model described in equation (4) on our data, separately by potential age of child (p). Panel (a) shows effects for women, panel (b) for partners. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time.

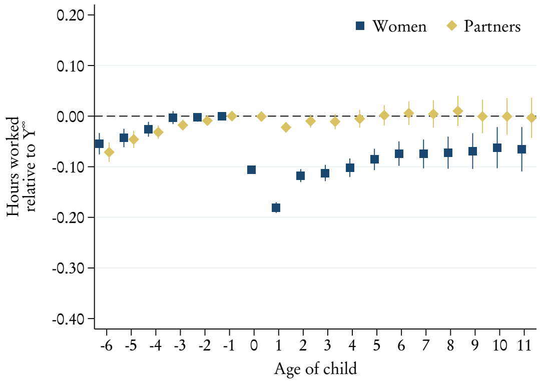


(a) Woman's earnings

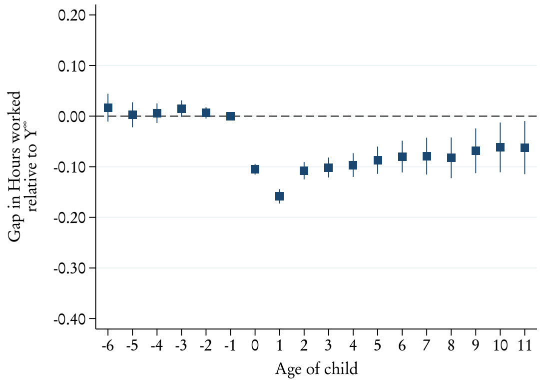
(b) Partner's earnings

Figure A2. Results by increasing horizon of potential ages p of child

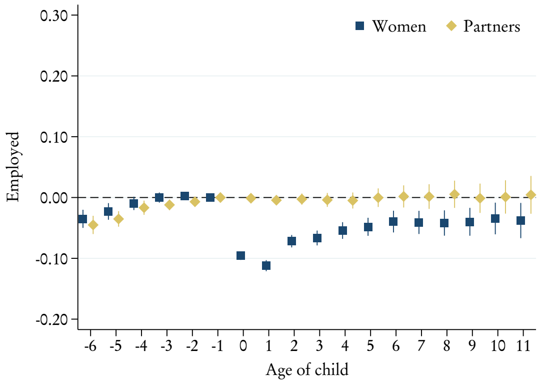
Note: Estimated effects of age of child on earnings using the event-IV model described in equation (10, event-IV), estimated by increasing horizon of potential ages a. Panel (a) shows effects for women, panel (b) for partners. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time.



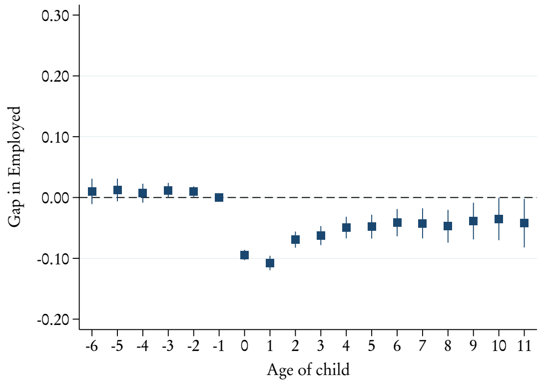
(a) Hours worked



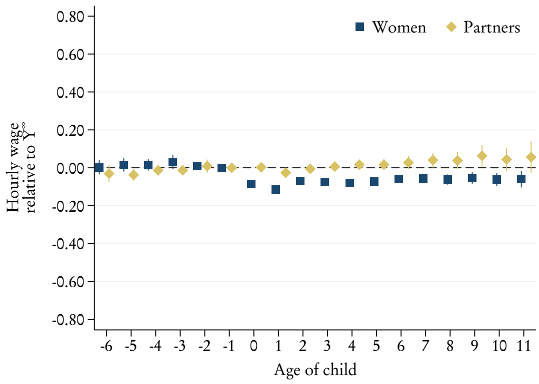
(b) Gap in hours worked



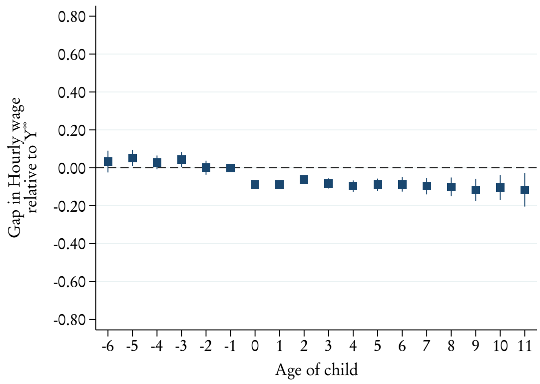
(c) Employment



(d) Gap in employment



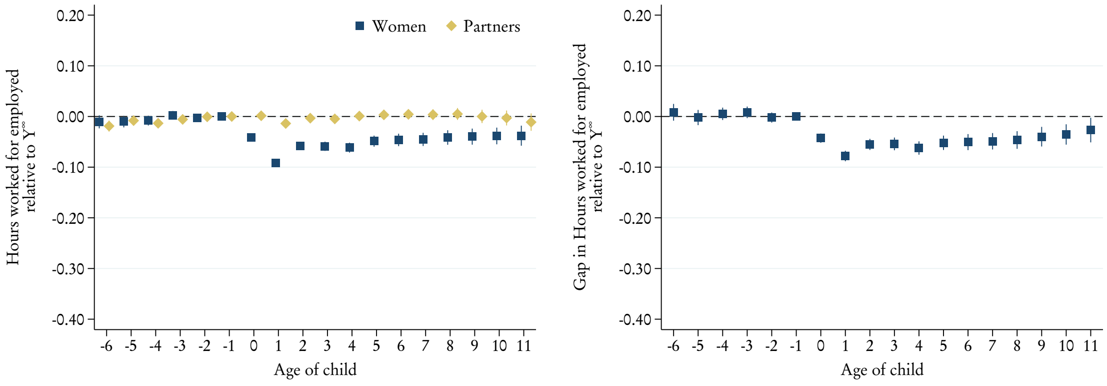
(e) Hourly earnings



(f) Gap in hourly earnings

Figure A3. Other labor market outcomes. Event.

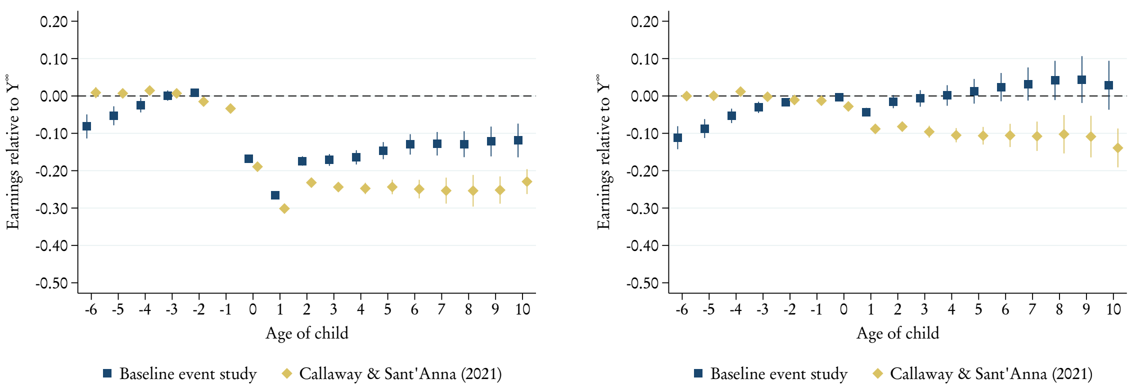
Note: Event-study estimates from specification (2). Outcomes are hours worked (panel a and b), employment (panel c and d), and hourly wages (panel e and f). Panel a, c, and e show effects separately for women and partners, figures b, d, and f show difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 3.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were registered with a partner at the time, and eventually had at least one child (observations = 145,571; women = 8,349).



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

Figure A4. Hours conditional on employment. Event.

Note: Event-study estimates from specification (2). Outcome is hours worked conditional on employment. Panel a shows effects separately for women and partners, panel b shows difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 3.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were registered with a partner at the time, and eventually had at least one child (observations = 145,571; women = 8,349).



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

Figure A5. Results based on the treatment-cohort Callaway and Sant'Anna (2021) estimator

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

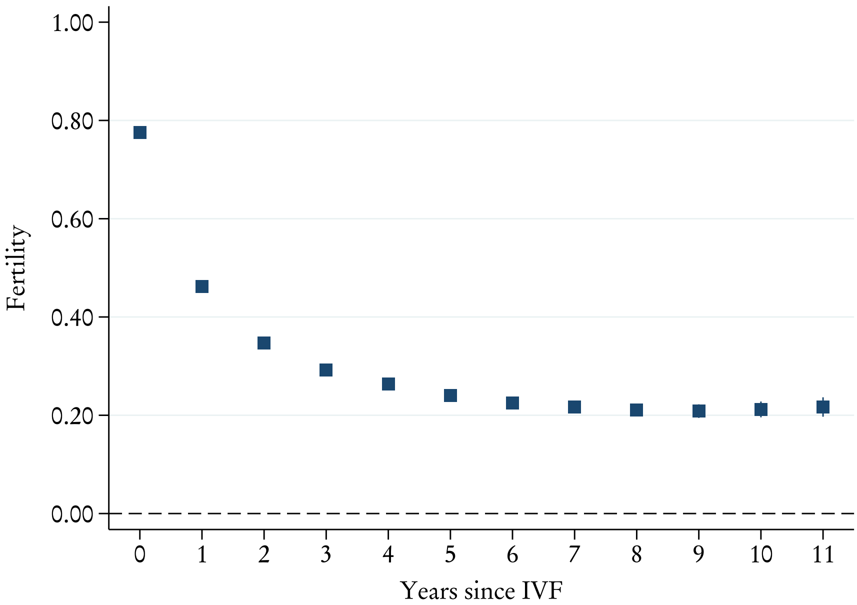
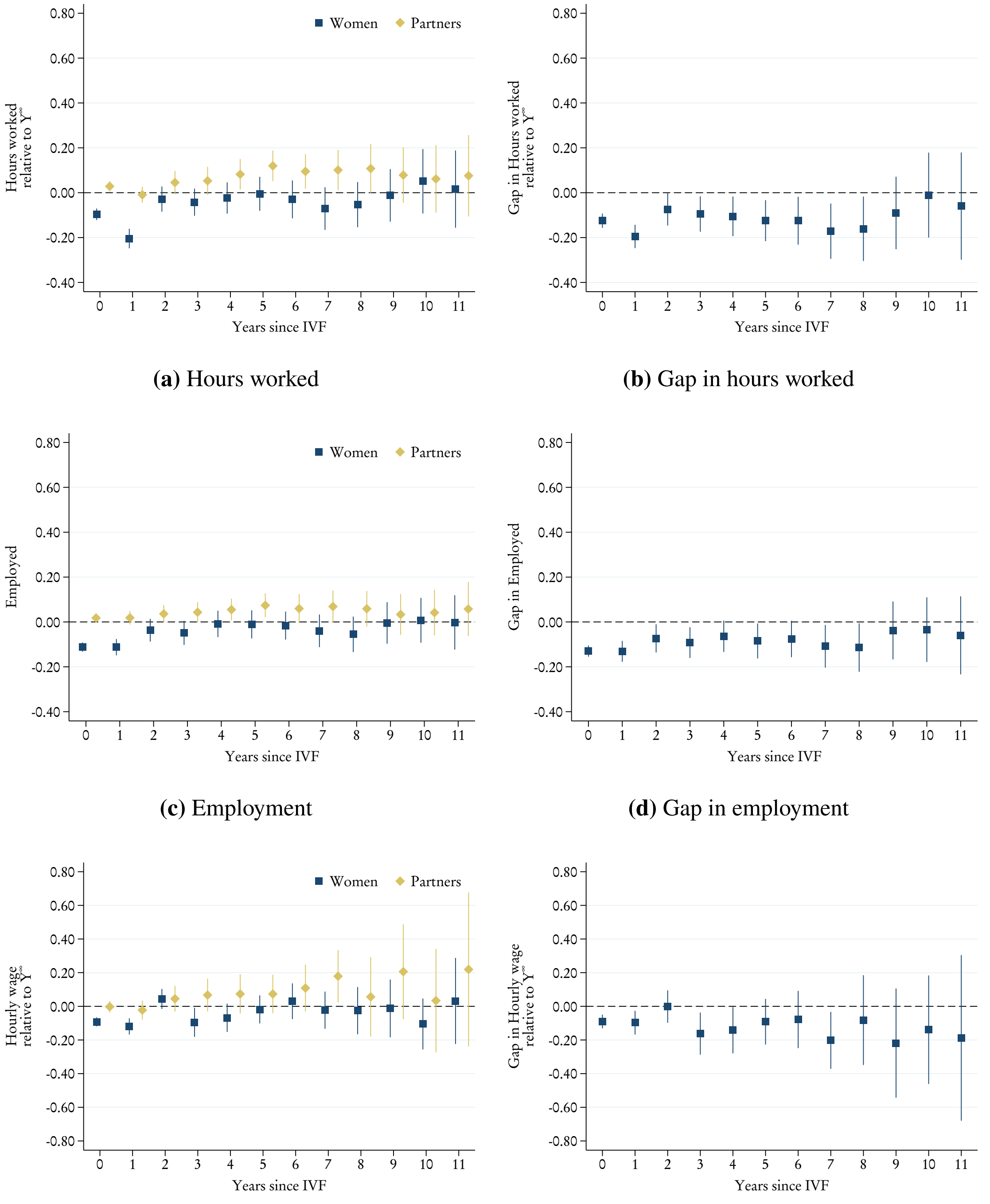


Figure A6. First stage. LPR-IV.

Note: First stage estimates using the IV model of Lundborg et al. (2017) as described in equation (5) on our data. The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; women = 10,033)



(a) Hours worked

(b) Gap in hours worked

(c) Employment

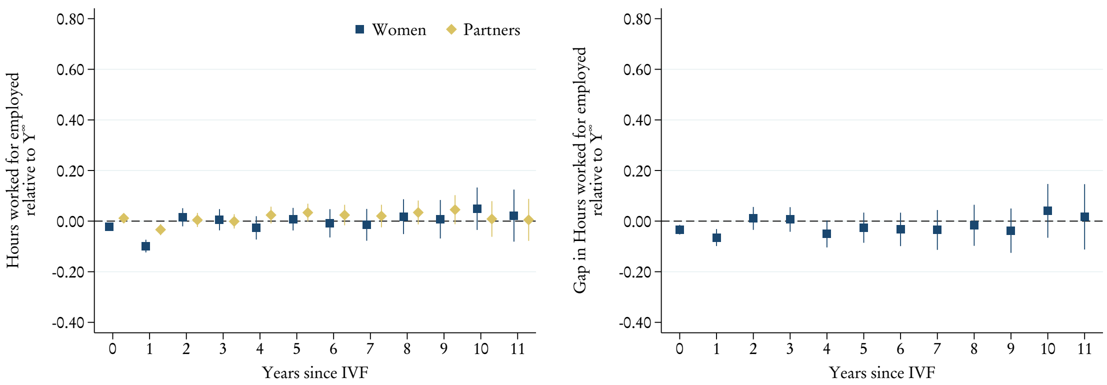
(d) Gap in employment

(e) Hourly earnings

(f) Gap in hourly earnings

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

Note: Estimated effects of fertility using the IV model of Lundborg et al. (2017) as described in equation (4) on our data. Outcomes are hours worked (panel a and b), employment (panel c and d), and hourly wages (panel e and f). Panels a, c, and e show effects separately for women and partners, panels b, d, and f show difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 3.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; women = 10,033)

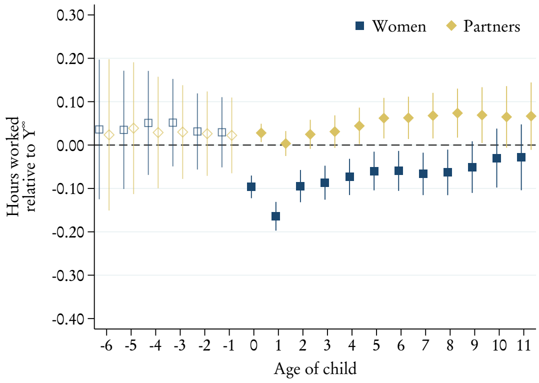


(a) Hours worked for employed

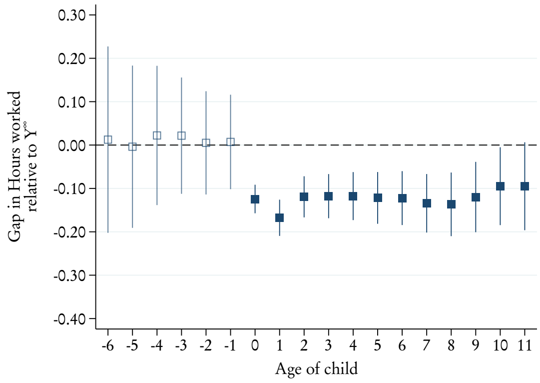
(b) Gap in hours worked for employed

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

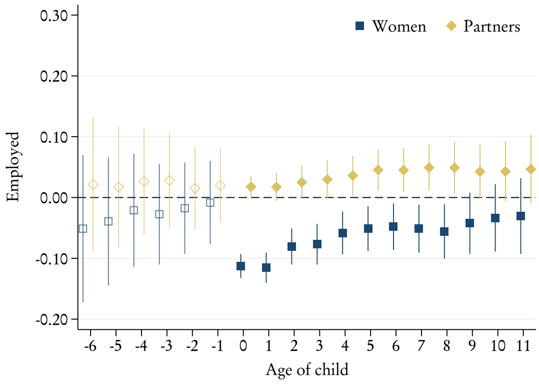
Note: Estimated effects of fertility using the IV model of Lundborg et al. (2017) as described in equation (4) on our data. Outcome is hours worked conditional on employment. Panel a shows effects separately for women and partners, panel b shows difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 3.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; women = 10,033).



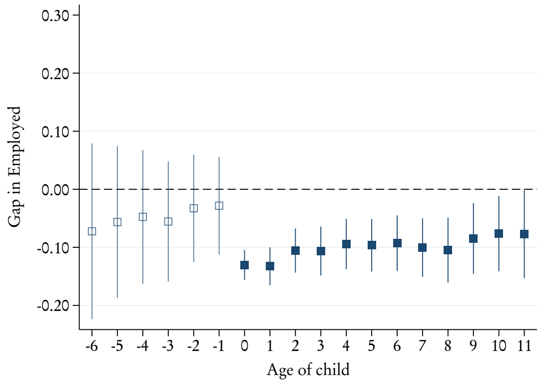
(a) Hours worked



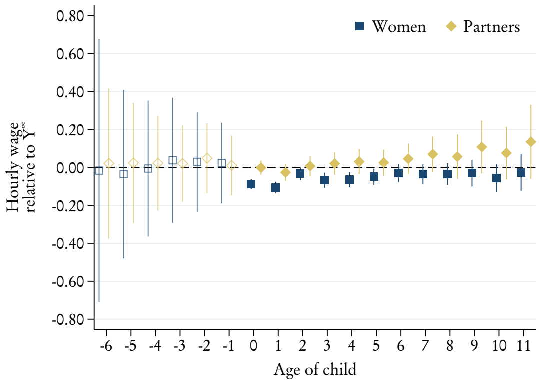
(b) Gap in hours worked



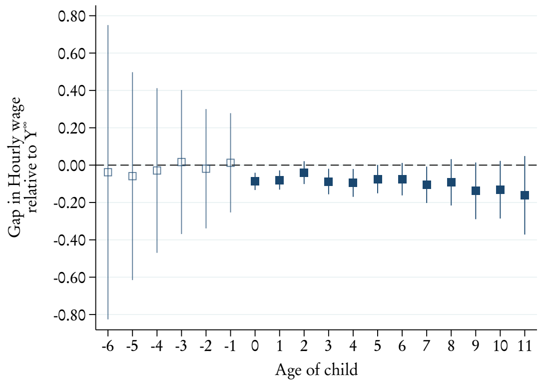
(c) Employment



(d) Gap in employment



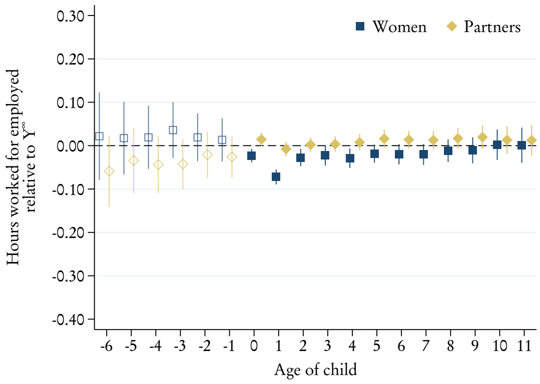
(e) Hourly earnings



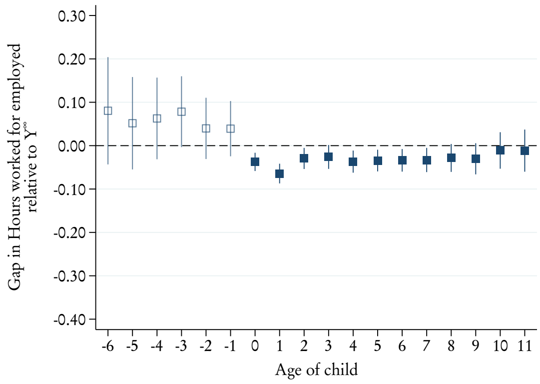
(f) Gap in hourly earnings

Figure A9. Other labor market outcomes. Event-IV.

Note: Estimated effects of age of child using the event-IV model described in equation (10, event-IV). Outcomes are hours worked (panel a and b), employment (panel c and d), and hourly wages (panel e and f). Panels a, c, and e show effects separately for women and partners, panels b, d, and f show difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 3.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; women = 10,033)



(a) Hours worked for employed



(b) Gap in hours worked for employed

Figure A10. Hours conditional on employment. Event-IV.

Note: Estimated effects of age of child using the event-IV model described in equation (10, event-IV). Outcome is hours worked conditional on employment. Panel a shows effects separately for women and partners, panel b shows difference between women and partners. Estimates are scaled relative to counterfactual earnings without children (Y^{∞}) as described in section 3.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; women = 10,033)

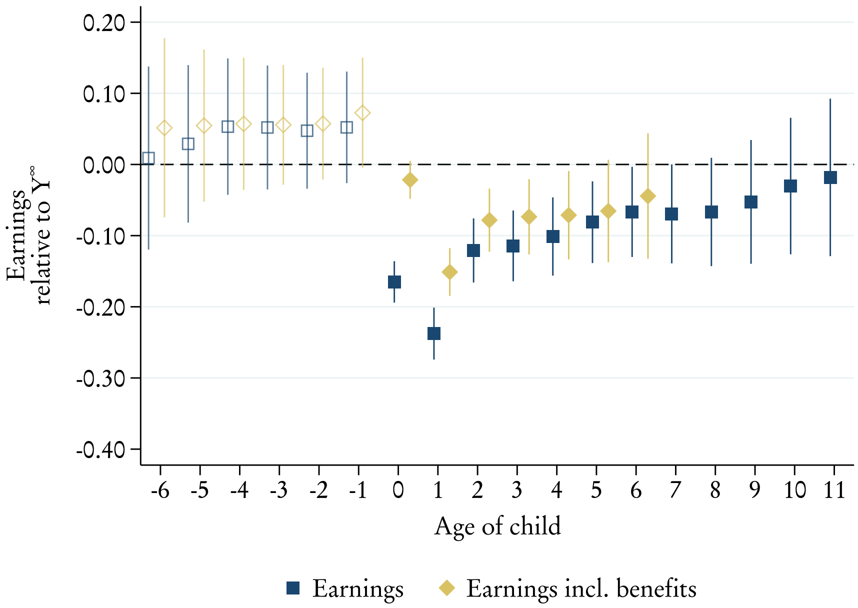
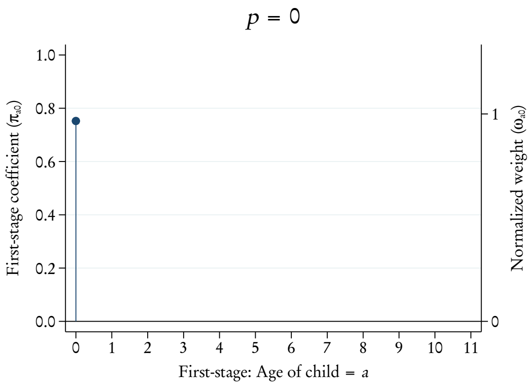
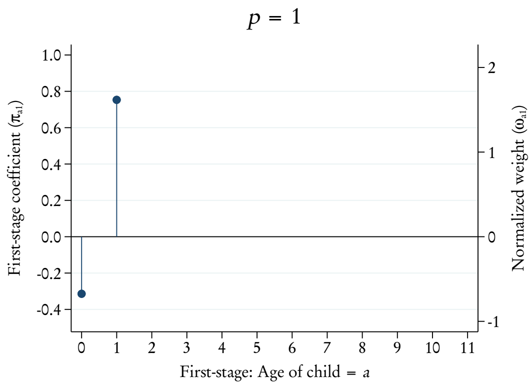


Figure A11. Earnings including benefits. Event-IV.

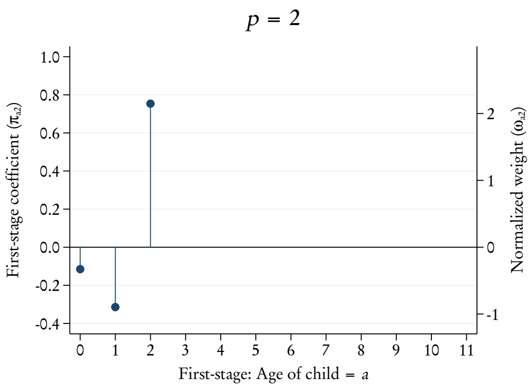
Note: Estimates from our Event-IV model as specified in equation (10, event-IV). Outcomes are earnings and earnings including benefits. Data on earnings including benefits are only available until 2017. Estimates are scaled relative to counterfactual earnings (Y^{∞}) as described in section 3.5. The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; women = 10,033)



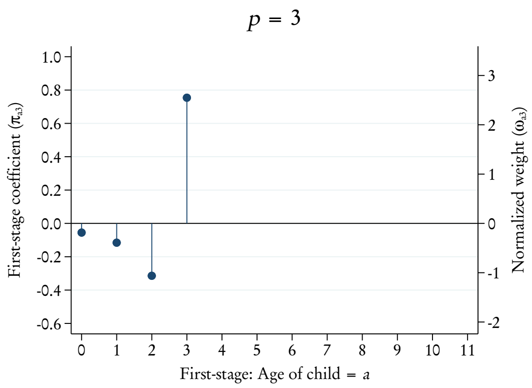
(a) Fertility weight at $p = 0$



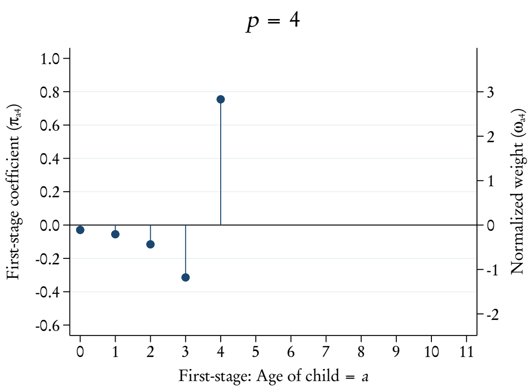
(b) Fertility weight at $p = 1$



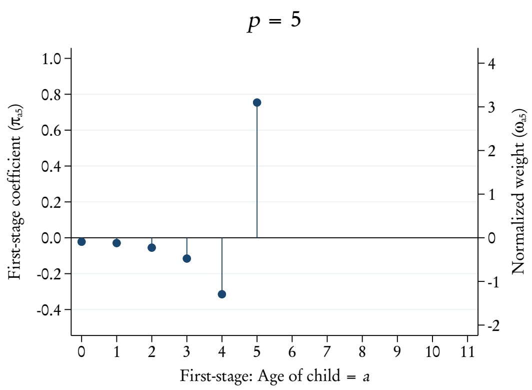
(c) Fertility weight at $p = 2$



(d) Fertility weight at $p = 3$



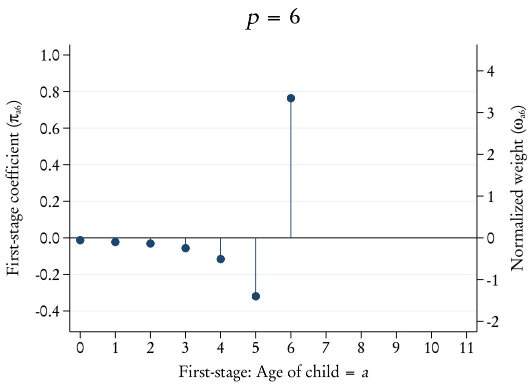
(e) Fertility weight at $p = 4$



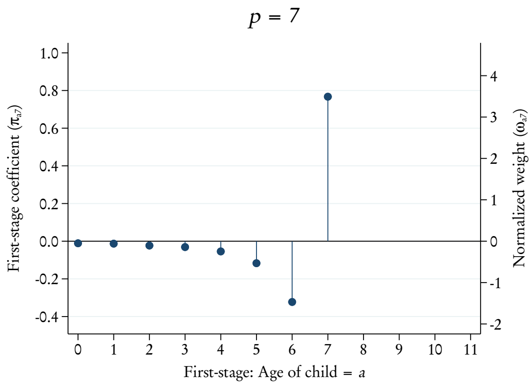
(f) Fertility weight at $p = 5$

Figure A12. Fertility weights

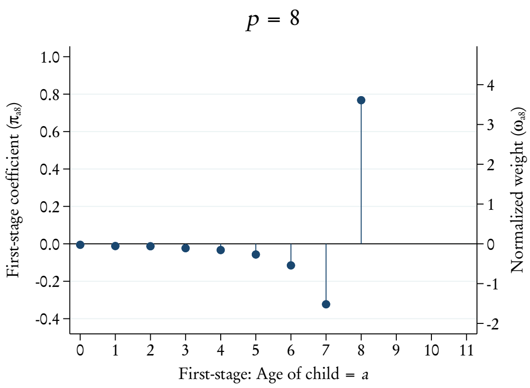
Note: Fertility weights as defined in Section 6.1. The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; women = 10,033).



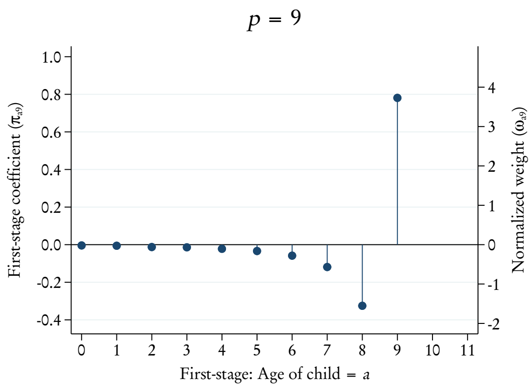
(a) Fertility weight at $p = 6$



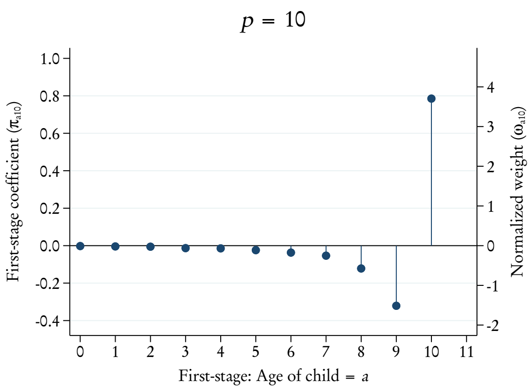
(b) Fertility weight at $p = 7$



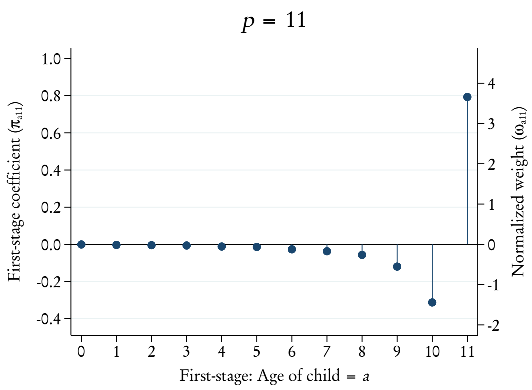
(c) Fertility weight at $p = 8$



(d) Fertility weight at $p = 9$



(e) Fertility weight at $p = 10$



(f) Fertility weight at $p = 11$

Figure A13. Fertility weights

Note: Fertility weights as defined in Section 6.1. The sample includes all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; women = 10,033).

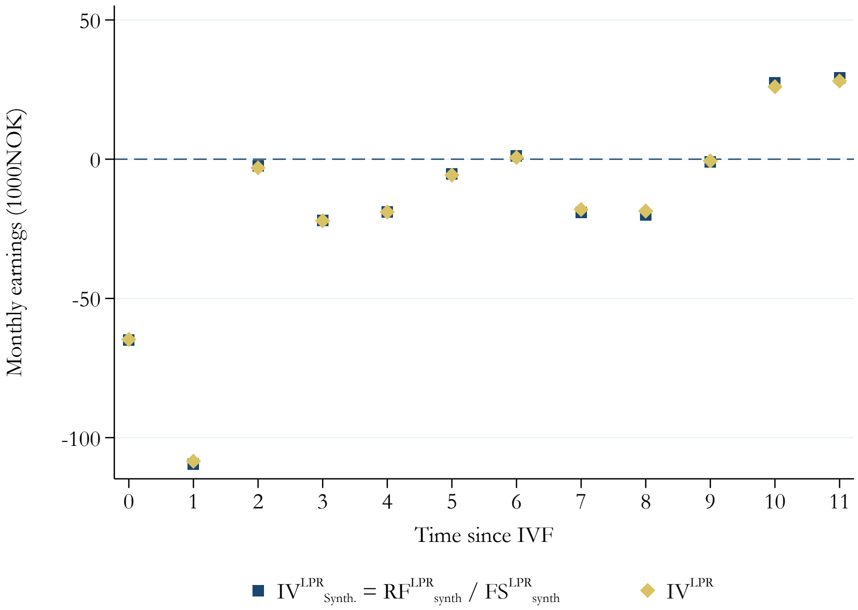
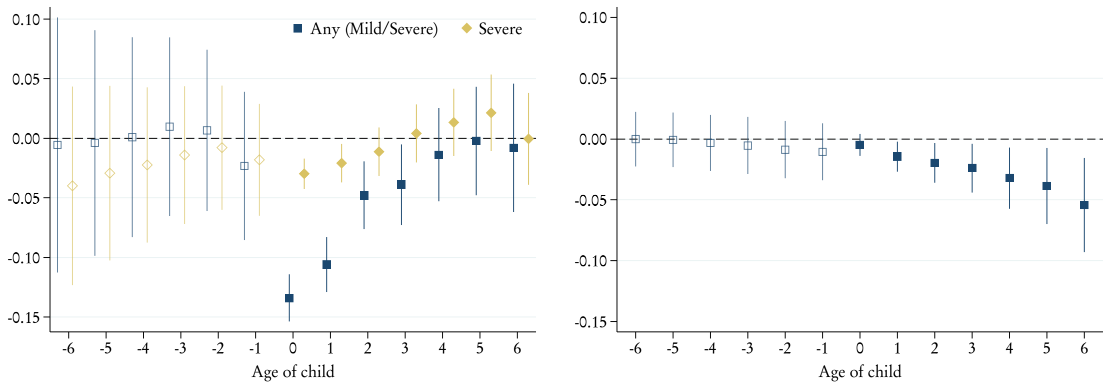


Figure A14. Combining results from LPR-IV and our event-IV model

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



(a) Psychological symptoms

(b) Divorce

Figure A15. Non-labor market outcomes – Event-IV.

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

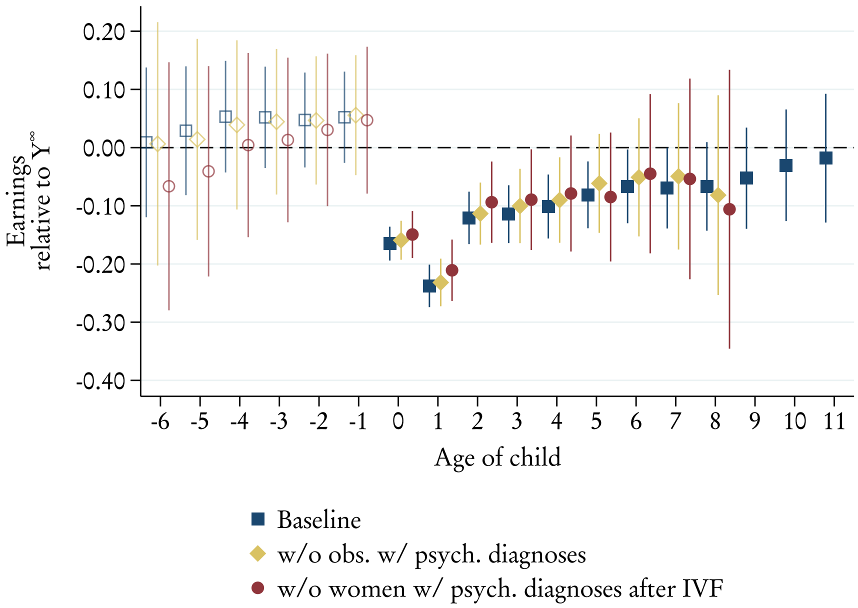
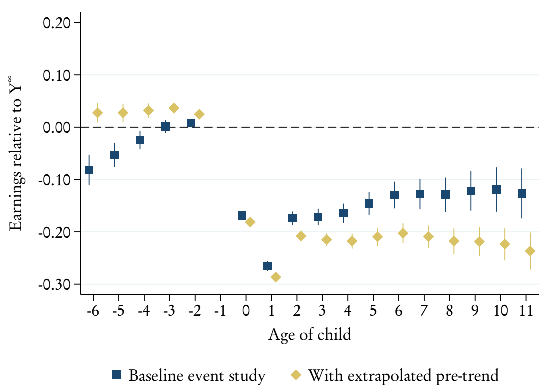
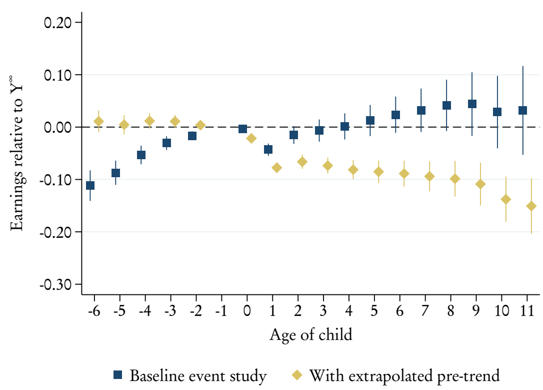


Figure A16. Robustness. Event-IV.

Note: Robustness checks of our event-IV model as specified in equation (10, event-IV). Figure A16 show our baseline specification estimated in the sample of IVF women (all women who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, and were registered with a partner at the time (observations = 173,480; women = 10,033)), alongside estimates based on (i) a subsample where we exclude all observations with a psychological diagnosis and (ii) a subsample where we exclude all *women* who received a psychological diagnosis after IVF treatment. All estimates are scaled relative to counterfactual earnings (Y^{∞}) as described in section 3.5.



(a) Woman's earnings



(b) Partner's earnings

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

Note: This Figure A17 reports event-study estimates that adjust for the baseline of a linear extrapolation of the pre-trend into the post period following Rambachan and Roth (2023). Panel (a) shows estimates for women, panel (b) shows the estimates for partners. Estimates are scaled relative to each gender's counterfactual earnings (Y^{∞}), as described in section 3.5. The sample includes all women (and their partners) who underwent their first IVF treatment between 2009 and 2016, had no children prior to that attempt, were registered with a partner at the time, and eventually had at least one child (observations = 145,571; women = 8,349).