Cohort-Chained DiD: Long-Run Effects with Limited Pre-Treatment Data

Dylan Balla-Elliott Isaac Norwich*

September 2025

Preliminary; latest version available here

Heterogeneity robust difference-in-differences methods typically require control units that remain untreated throughout the entire post-treatment window. This prevents the identification of long-run effects when researchers observe fewer pre-treatment periods than post-treatment periods. We show that cohort-stacked estimators identify long-run effects by chaining together successive not-yet-treated controls. This approach uses overlapping cohorts to extend identification under standard common trends assumptions. We apply this to the earnings effects of parenthood. In a setting where direct methods identify effects only four years post-birth, chaining extends identification to eight years.

*Balla-Elliott: University of Chicago, dbe@uchicago.edu. Norwich: University of Chicago, inorwich@uchicago.edu.
We thank Kory Kroft, Matthew Notowidigdo, and Stephen Tino for helpful comments. Our presentation of the canonical DiD arguments and the notation throughout are heavily influenced by Alex Torgovitsky's course materials, though all errors are our own. This material is based on work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1746045. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Direct difference-in-differences methods recover only a subset of the identified dynamic treatment effects when researchers observe fewer pre-treatment periods than post-treatment periods. These methods directly compare changes in treated units to changes in control units (Callaway and Sant'Anna, 2021). This requires control units that remain untreated throughout the entire observation window for each treated cohort. When researchers observe fewer pre-treatment periods than post-treatment periods, these methods cannot identify long-run effects because no control cohort spans the full window.

This limitation is common in practice. Administrative datasets often contain units for only a few periods before treatment. Alternatively, researchers may deliberately restrict controls to use only soon-to-be-treated units.

We solve this problem by chaining successive control cohorts. The cohort-chained difference-in-differences (CCDID) estimator reconstructs counterfactual trends through intermediate periods using overlapping cohorts.[1] If cohort A provides the trend from periods 1 to 3, and cohort B provides the trend from periods 3 to 5, chaining recovers the full trend from periods 1 to 5. This approach requires only that standard common trends and no anticipation assumptions hold across all cohorts and time periods.

Under the standard assumptions, direct comparisons unnecessarily restrict the control group and recover only a subset of the parameters identified by the assumptions and data. When common trends hold, the rank condition necessary for chaining is weaker than requiring the same control cohorts in both reference and target periods. Overlapping cohorts provide sufficient variation to identify period fixed effects that connect distant time periods.[2]

We use this approach to estimate the effects of parenthood on earnings. Following Cortés and Pan (2023), we restrict to parents within five years of first birth. Standard methods identify effects only four years post-birth due to the five-year observation window. Chaining extends identification to eight years post-birth.

This note proceeds as follows. Section 1 formalizes the setting and assumptions and introduces our chaining identification strategy. Section 2 presents our regression-based estimation strategy using a cohort-stacked framework that jointly recovers all event-time treatment effects. Section 3 demonstrates the methodology through an application to the motherhood earnings penalty literature as in Cortés and Pan (2023). Section 4 concludes by discussing the implications of the chaining methodology for empirical practice in staggered adoption settings.

[1] It is also robust to heterogeneous treatment effects when there are multiple (staggered) treatment dates, like the recent estimators proposed by Sun and Abraham (2020), de Chaisemartin and D'Haultfœuille (2020), Callaway and Sant'Anna (2021), and Borusyak, Jaravel and Spiess (2024).

[2] Bellégo, Benatia and Dortet-Bernadet (2025) make a similar argument for unbalanced panels with attrition. We focus specifically on staggered treatment settings where cohorts are observed only for limited pre-treatment observation windows.

1 Identification: Chaining Cohorts with Common Trends

The cohort-chained estimator uses the same common trends and no anticipation assumptions as the "direct" estimators, but identifies dynamic effects up to longer time horizons. It relies on the easily testable condition that multiple not-yet-treated cohorts overlap and can be chained together.

1.1 The DiD Setting

All units receive an absorbing treatment, but there is variation in treatment timing.[3] Let G_i denote unit i's treatment period, Y_{it} the outcome, and $r_{it} = t - G_i$ the periods since treatment. Following the literature, let $Y_{it}(g)$ be the potential outcome in time t when treated in time g and $Y_{it}(\infty)$ be the untreated potential outcome.

The average treatment effect for cohort g in period t is:

$$\text{ATT}_t(g) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty)|G_i = g] \tag{1}$$

We observe the potential outcomes that correspond to the realized treatment timing $Y_{it} \equiv Y_{it}(G_i)$ not the counterfactual untreated potential outcome $\mathbb{E}[Y_{it}(\infty)|G_i = g]$. The untreated potential outcome is identified under the canonical common trends and no anticipation assumptions.

Assumption 1 (CT)**.** *For every pair of cohorts g and g' and every pair of time periods $t \neq t'$:*

$$\mathbb{E}[Y_{it}(\infty) - Y_{it'}(\infty)|G_i = g] = \mathbb{E}[Y_{it}(\infty) - Y_{it'}(\infty)|G_i = g'] \tag{2}$$

Assumption 2 (NA)**.** *For every cohort g and all $t < g$:*

$$\mathbb{E}[Y_{it}(g) \mid G_i = g] = \mathbb{E}[Y_{it}(\infty) \mid G_i = g] \tag{3}$$

1.2 The (Potentially Infeasible) Direct Approach

The direct DiD approach uses a not-yet-treated cohort $g' > t$ observed in both a reference period $s < g$ and target period t:

$$\mathbb{E}[Y_{it}(\infty)|G_i = g] = \underbrace{\mathbb{E}[Y_{is}|G_i = g]}_{\text{"level" from } s \text{ (NA)}} + \underbrace{\mathbb{E}[Y_{it} - Y_{is}|G_i = g']}_{\text{"trend" from } g' \text{ (CT)}} \tag{4}$$

The direct plug-in estimator is feasible when g' is observed in both s and t. When no such cohort exists (i.e. when there are fewer pre-treatment periods than post-treatment periods) the direct estimators following Callaway and Sant'Anna (2021) are not feasible.

[3] Either there are no never-treated units in the data, or researchers make a common trends assumption only among units that are ever treated.

1.3 Chaining Through Intermediate Periods

Suppose no single control cohort is observed in time periods 1 and 3. But, cohort A is present in periods 1 and 2 and cohort B is present in periods 2 and 3. Since the change from 1 to 3 is simply the change from 1 to 2 plus the change from 2 to 3, use A to identify $1 \to 2$ and B to identify $2 \to 3$. Then, "chain" those together to recover the change from 1 to 3.

More generally, consider two cohorts g', g'' and two periods s, t. Suppose cohort g' exists in periods p and t, while cohort g'' exists in periods s and p, where $s < p < t$. Under common trends, the trend from s to t is identified by chaining together cohorts g' and g'':

$$\begin{aligned}\mathbb{E}[Y_{it}(\infty)|G_i = g] = \mathbb{E}[Y_{is}|G_i = g] \\ + \underbrace{\left(\mathbb{E}[Y_{it} - Y_{ip}|G_i = g']\right)}_{\text{Trend from } p \text{ to } t} + \underbrace{\left(\mathbb{E}[Y_{ip} - Y_{is}|G_i = g'']\right)}_{\text{Trend from } s \text{ to } p}\end{aligned} \tag{5}$$

The first bracketed term uses g' to link periods p and t. The second uses g'' to link periods s and p. Together, they construct the full counterfactual trend from s to t.

1.3.1 An Example of When Chaining Works

Figure 1 presents a view of the relative years that exist for each cohort (x-axis) and calendar year (y-axis) in a hypothetical example where each cohort is observed for 4 years before treatment.

In this example, the direct estimator is only feasible up to relative time 2. In relative time 3, there is no single cohort with pre-treatment observations in both relative periods -1 and 3 for the cohort of interest. However, a "chain" of overlapping control cohorts does connect these periods. The orange squares highlight one such chain that identifies the counterfactual time trend.

2 Estimation: Chaining Cohorts through a Stacked Regression

We implement the chaining estimator through a cohort-stacked regression, similar to Wing, Freedman and Hollingsworth (2024).

2.1 Regression specification for a single cohort

For each treated cohort g, construct a "slice" by keeping all observations from cohort g and pre-treatment observations from not-yet-treated cohorts $g' > g$.

Formally, we keep the following observations:

$$\underbrace{\{(i,t) : G_i = g\}}_{\text{Treatment}} \text{ and } \underbrace{\{(i,t) : G_i > g \text{ and } t < G_i\}}_{\text{Control}} \tag{6}$$

Chaining requires overlap between sequential cohorts to reconstruct counterfactual trends through intermediate periods.[4] This is a testable condition. Researchers can verify whether their data structure permits chaining for desired horizons.

Within this subset, estimate:

$$Y_{it} = \sum_{g'} \gamma_{g'} \mathbf{1}\{G_i = g'\} + \sum_{s} \tau_s \mathbf{1}\{t = s\} + \sum_{r \neq -1} \delta_{rg} \mathbf{1}\{G_i = g, r = t - g\} + \epsilon_{it} \tag{7}$$

where $\gamma_{g'}$ are cohort fixed effects, τ_s are period fixed effects, and δ_{rg} captures treatment effects for the cohort of interest g at relative time r.

This specification uses all available pre-treatment cohorts as controls and estimates all relative-time effects for cohort g simultaneously. The chaining argument identifies the period fixed effects τ_s, even as the composition of control cohorts changes. See Appendix A.1 for further discussion.

2.2 The stacked regression with many cohorts

Instead of estimating effects for each cohort separately, create an estimation dataset that "stacks" each cohort-specific "slice." Also, create a variable with the "slice" identifier. Then, fully interacting all the variables in the regression in Equation (7) with the vector of slice indicators recovers the same point estimates as the series of subset-specific regressions.

Let k denote the treated cohort in a particular slice. The regression is then fully interacted with the slice indicators :

$$Y_{itk} = \sum_{k} \sum_{g'} \gamma^k_{g'} \mathbf{1}\{G_i = g'\} + \sum_{k} \sum_{s} \tau^k_s \mathbf{1}\{t = s\} + \sum_{k} \sum_{r \neq -1} \delta_{rk} \mathbf{1}\{G_i = k, r = t - k\} + \epsilon_{itk} \tag{8}$$

Notice now that the cohort fixed effects $\gamma^k_{g'}$, the time trends τ^k_s, and the dynamic effects δ_{rk} are slice-specific and thus do not include undesired cross-slice contrasts. The stacked regression is useful since jointly estimating these parameters recovers the full variance-covariance matrix for all cohort-by-relative time estimates. This makes it straightforward to calculate the standard errors for linear combinations of the slice-specific treatment effects for cohort k in relative period r, δ_{rk}.

The slice-specific treatment effects δ_{rk} represent the treatment effect for cohort k at relative time r. One natural way to aggregate these into relative-time effects across cohorts is to weight by cohort size:

$$\text{ATT}_r = \sum_{g} \text{ATT}_r(g) \cdot \mathbb{P}[G_i = g | \text{observed at } r] \tag{9}$$

[4]The matrix of fixed effects has full rank if and only if all control cohorts and time periods belong to the same connected set. Dynamic effects for time periods outside of this connected set are not identified. With multiple treated cohorts and multiple slices, each slice has its own connected set of control cohorts and time periods that determine the set of feasible dynamic treatment effects.

This manual aggregation provides full control over weighting but in principle requires additional computation. However, modern regression software often provides routines to estimate an interacted model and then aggregate to a weighted average.[5] In Appendix A.3, we suggest a regression-based routine that can be used to aggregate over slices in cases where computational limits make it impossible to estimate the full vector of cohort-by-relative time effects.[6]

3 Application: The Effects of Children on Parental Earnings

Event studies are often used to study the effect of the arrival of children on labor market earnings. In this literature, it is common to have fewer pre-treatment than post-treatment observations. For example, Cortés and Pan (2023) use 5 years of pre-treatment data and are interested in long-run effects up to 8 years after treatment. Kleven, Landais and Søgaard (2019) also use 5 years of pre-treatment data and is interested in long-run effects up to 10 years after treatment. These longer run effects are identified in the canonical TWFE specification, which can include unintended contrasts with already-treated units (Roth et al., 2023).

3.1 Setting

Cortés and Pan (2023) use the 1976 to 2017 waves of the Panel Study of Income Dynamics (PSID) to estimate the impacts of parenthood on earnings. They restrict the PSID sample to household heads and spouses/cohabiters between the ages of 20 and 55 years old and who had their first child between the ages of 20 and 45. Further inclusion criteria include parents who are observed at least once before and after the birth of their first child and whose earnings outcomes are observed at least four times during the fifteen-year window (five periods before and 10 periods after) surrounding the year of birth. The main outcome in their paper is annual labor earnings (total labor income before taxes and transfers for the year prior to the interview).

3.2 Implementation

We follow Cortés and Pan (2023) and restrict to parents within 5 years of the birth of their first child in the main specification. We also interact the time trend with the parent's year of birth to allow for life-cycle effects. Effects are estimated separately for men and women. The thought experiment is thus to compare women born in 1990 who first become mothers in 2020 to other women born in 1990 who first become mothers between 2021 and 2025. Aggregate estimates are then an average over birth years of the parents (e.g. 1990) and year of birth of the first child (e.g. 2020).

[5] A leading example is the `fixest` package, which has a simple `sunab` command that will estimate the full set of $\text{ATT}_r(g)$ and aggregate.

[6] This procedure requires careful re-weighting before estimation to ensure that the regression aggregator has desired weights. Additionally, only one dynamic time effect can be estimated in each regression, requiring as many regressions as there are dynamic effects of interest. However, this procedure has the advantage of estimating the average dynamic effects in a single step, without estimating the high dimensional covariance matrix for all of the cohort-by-relative time effects. This procedure is computationally attractive in very large datasets with very many cohorts.

We estimate long-run effects in the spirit of their Figure 1. Like Cortés and Pan (2023), we report the earnings as shares of the pre-period mean. We also exclude periods after 1997 to avoid gaps in the data.[7]

3.3 Results

Figure 2a reports heterogeneity-robust estimates using the CSA estimator. Due to the sample restriction that control units must be within five years of the birth of their first child, we can only estimate the first four post-event relative time coefficients. Figure 2b reports heterogeneity-robust estimates using our cohort-stacked estimator, in which the full set of eight relative time coefficients are identified. The benefit of our estimator is that we recover the time-path up to eight years post-event, at which effects attenuate and stabilize at a reduction of around $10,000 in earnings.

Figure 3 then highlights that long-run effects are still identified using as few as two years of data before the arrival of the first child. In each case, we are able to identify long-run (i.e. eight-year) effects.

4 Conclusion

This note highlights that the regression-based CCDID estimator can identify long run effects that are not identified in direct estimators. The application to Cortés and Pan (2023) highlights the practical value for empirical researchers working with limited pre-treatment data.

We show that a carefully constructed regression implicitly "chains" together control cohorts to expand the set of identifiable treatment effects, relative to what direct DiD estimators recover. By chaining common trends assumptions across overlapping cohorts in calendar time, the regression imputes counterfactual trends through intermediate periods, enabling identification of treatment effects at longer horizons than direct comparisons allow.

[7]The PSID collected data annually until 1997, and starting in 1999 collected data biennially.

References

Abowd, John M., Francis Kramarz, and David N. Margolis. 1999. "High Wage Workers and High Wage Firms." *Econometrica*, 67(2): 251–333.

Bellégo, Christophe, David Benatia, and Vincent Dortet-Bernadet. 2025. "The Chained Difference-in-Differences." *Journal of Econometrics*, 248: 105783.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting Event Study Designs: Robust and Efficient Estimation." *Review of Economic Studies*.

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics*, 225(2): 200–230.

Cortés, Patricia, and Jessica Pan. 2023. "Children and the Remaining Gender Gaps in the Labor Market." *Journal of Economic Literature*, 61(4): 1359–1409.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review*, 110(9): 2964–2996.

Kleven, Henrik, Camille Landais, and Jakob Egholt Søgaard. 2019. "Children and Gender Inequality: Evidence from Denmark." *American Economic Journal: Applied Economics*, 11(4): 181–209.

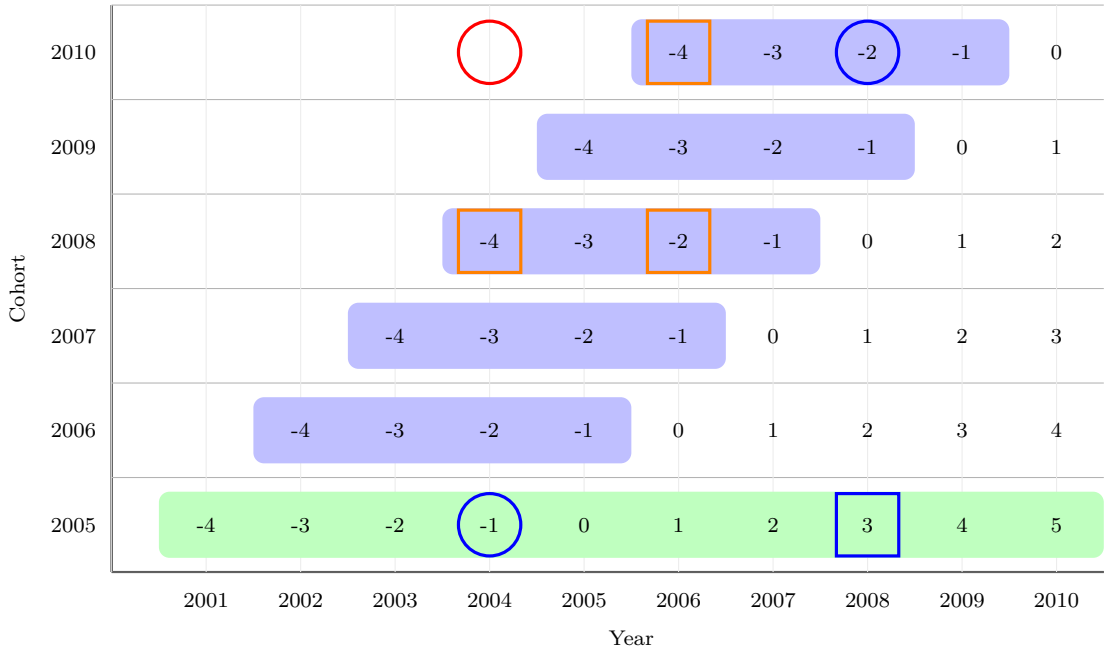
Roth, Jonathan, Pedro H.C. Sant'Anna, Alyssa Bilinski, and John Poe. 2023. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature." *Journal of Econometrics*, 235(2): 2218–2244.

Sun, Liyang, and Sarah Abraham. 2020. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics*.

Wing, Coady, Seth Freedman, and Alex Hollingsworth. 2024. "Stacked Difference-in-Differences." *NBER WP*.

Figures

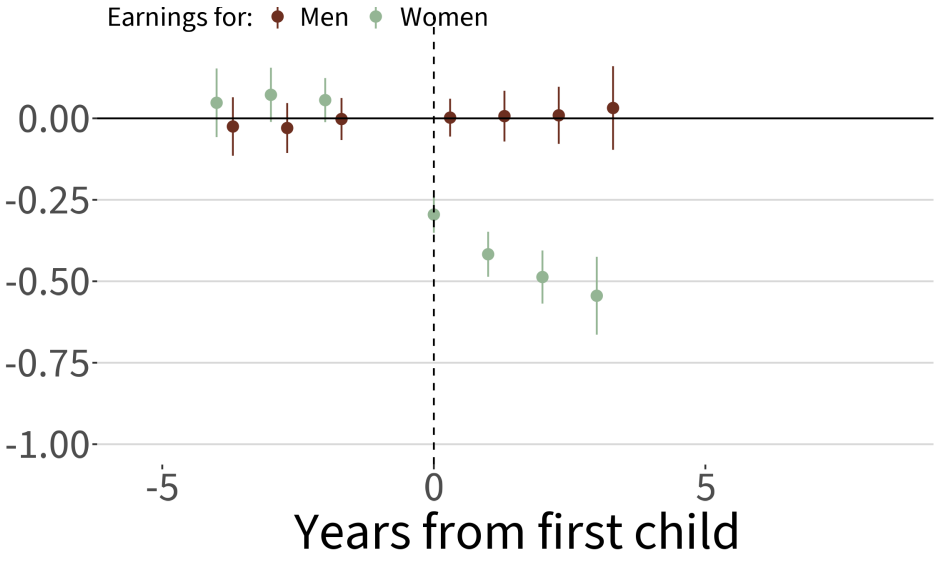
Figure 1: Example Panel Setting



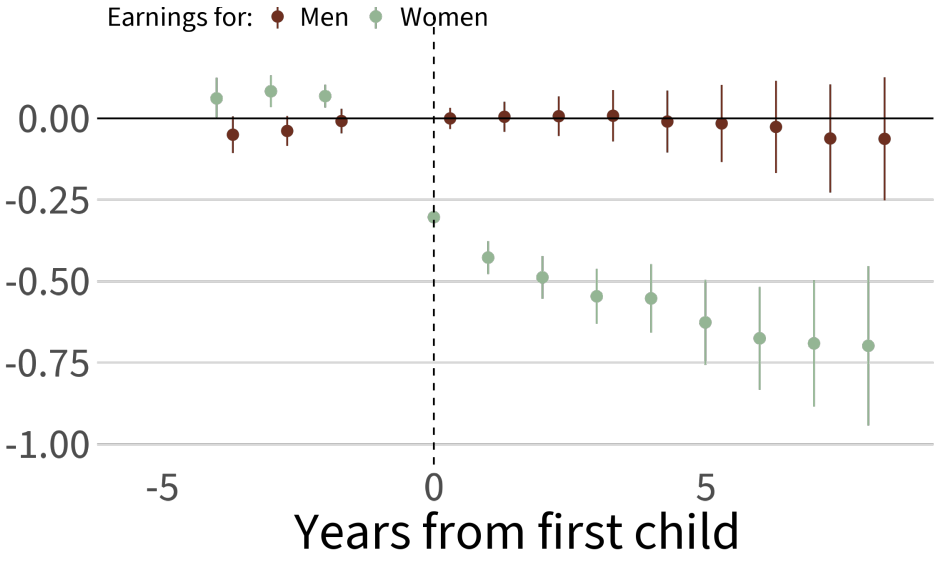
Notes: This figure shows an example panel setting with $r_{\min} = -4$ and $T_{\max} = 2010$ for cohorts $g \in \mathcal{G} = \{2005, \ldots, 2010\}$. The green shading denotes the observations for the cohort of interest, $g = 2005$, while the purple shading highlights later-treated cohorts' pre-treatment relative years. In the direct DiD estimator, the difference between $(t, g) = (2008, 2005)$ to $(2004, 2005)$ and $(2008, 2010)$ to $(2004, 2010)$ would identify the target parameter $\text{ATT}_{2008}(2005)$ (blue square). As noted by the red circle, the observation at $(2004, 2010)$ is missing in our example since $r_{\min} = -4$. Thus there is no individual cohort that acts as a control group, as is necessary in direct DiD estimators (Callaway and Sant'Anna, 2021). However, as we show in this note, we can use the trend from the cohort with $g'' = 2008$ under Assumption 1 (CT) to chain together two trends. The orange squares represent the additional observations used to identify this counterfactual time trend through chaining.

Figure 2: Stacked DiD Identifies Long-Run Effects

(a) CSA DiD Estimates: Within 5 Years of First Child

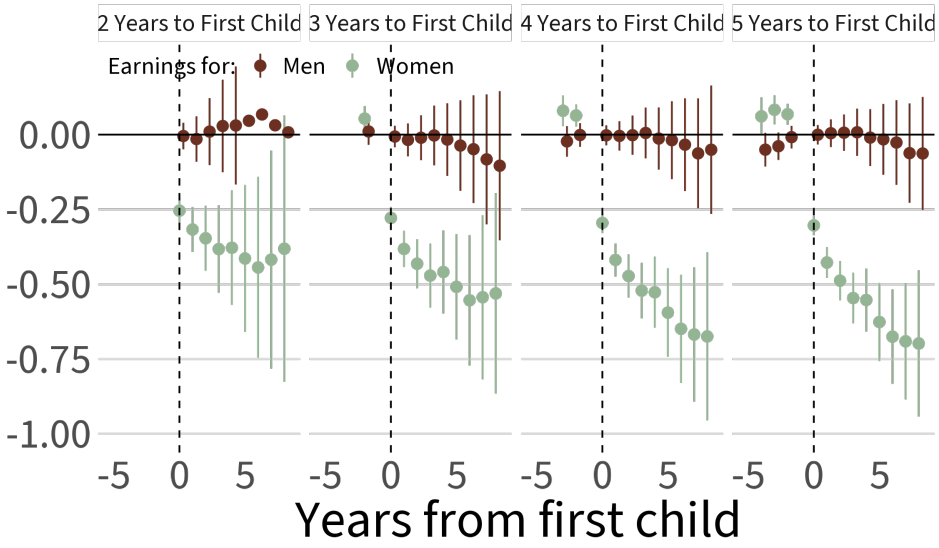


(b) Stacked DiD Estimates: Within 5 Years of First Child



Notes: This figure presents estimates of parenthood on labor earnings for men and women. We adopt the sample restrictions from Cortés and Pan (2023), where we keep only household heads and spouses/cohabiters between the ages of 20 and 55 years old who had their first child between the ages of 20 and 45. We keep parents who are observed at least once before and after the birth of their first child and whose earnings outcomes are observed at least four times during the fifteen-year window (five periods before and 10 periods after) surrounding the year of birth. Panel (a) reports the event study using the estimator from Callaway and Sant'Anna (2021). Due to the 5-year pre-treatment window, these results are only estimated up to relative year 4. Panel (b) reports the event study using the CCDID estimator presented in this paper. With the same pre-treatment restriction, we are able to estimate effects up to 8 years post-treatment.

Figure 3: Stacked DiD Estimates: Various Years to First Child



Notes: This figure presents estimates of parenthood on labor earnings for men and women using the CCDID estimator. We adopt the sample restrictions from Cortés and Pan (2023), where we keep only household heads and spouses/cohabiters between the ages of 20 and 55 years old who had their first child between the ages of 20 and 45. We keep parents who are observed at least once before and after the birth of their first child and whose earnings outcomes are observed at least four times during the fifteen-year window (five periods before and 10 periods after) surrounding the year of birth. We vary the pre-treatment window to include individuals within two, three, four, or five years of their first child. We are thus able to identify long-run effects in these samples.

Appendix

A Identification

A.1 Recovering cohort-specific treatment effects from within-slice regression

A.1.1 Standard DiD identification within slice

Consider the regression for treated cohort g:

$$Y_{it} = \sum_{g'} \gamma_{g'} \mathbf{1}\{G_i = g'\} + \sum_{s} \tau_s \mathbf{1}\{t = s\} + \sum_{r \neq -1} \delta_{rg} \mathbf{1}\{G_i = g, r = t - g\} + \epsilon_{it} \tag{10}$$

Under common trends and no anticipation, this regression is correctly specified. The treatment effect is:

$$\text{ATT}_r(g) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty) | G_i = g] = \delta_{rg} \tag{11}$$

This is the basic DiD result without staggered timing. The δ_{rg} coefficients are the difference between treated outcomes and the untreated potential outcomes implied by the time trends and cohort fixed effects.

A.1.2 Identification under partial overlap

Unlike direct DiD estimators that require complete overlap between treatment and control cohorts, this regression remains identified when cohorts have only partial overlap. The rank condition requires only that there is sufficient overlap between control cohorts to build a "chain" that connects the reference time period to the time periods of interest. The matrix of fixed effects has full rank if and only if all control cohorts and time periods in every slice belong to the same connected set. [8]

When this condition holds, the design matrix has full rank, which enables the separate identification of $\gamma_{g'}$, τ_s, and δ_{rg}. This overlap requirement is simple to check in practice.

A.2 Why One-Shot Aggregation of All Relative Times Jointly Fails

We form a stacked dataset where each observation is indexed by (i, t, k), with k denoting the slice, i the unit, and t the calendar period. Consider the short regression of Y_{it}^k on $\widetilde{D}_{it}^k(r)$ which is the residualized indicator for relative time r. The long regression includes slice-by-cohort, slice-by-time, and the full vector of relative time indicators $D_{it}^k(r) \equiv \mathbb{1}[G_i = k, t = k + r]$, except for $D_{it}^k(-1)$,

[8] This follows the logic of Abowd, Kramarz and Margolis (1999) for identifying worker and firm effects. Instead of forming a bipartite graph of workers and firms, form a bipartite graph of control cohorts and time periods, with an edge connecting cohort g with time t if g is observed in t. When this fails, there is a set of time period indicators and a cohort indicator that are perfectly collinear and therefore not separately identified. Repeat this test for every slice; the connected sets of control cohorts will vary across slices.

and a constant. Thus $\widetilde{D}^k_{it}(r)$ is obtained by regressing $D^k_{it}(r)$ on slice-by-cohort and slice-by-time indicators and taking the residual.

Appealing to FWL, we can write the coefficient in the long regression using the short regression.

$$\hat{\delta}_r \sum_k \sum_t \sum_i (\widetilde{D}^k_{it}(r))^2 = \sum_k \sum_t \sum_i Y^k_{it} \widetilde{D}^k_{it}(r) \tag{12}$$

Parallel trends implies that the Y^k_{it} have an additively separable form with $\mathbb{E}[\epsilon_{it} \mid G_i] = 0$ for all t.

$$Y_{it} = \alpha_g + \gamma_t + \sum_{r \geq 0} \delta_{rg} D_{it}(r) + \epsilon_{it} \tag{13}$$

Substitute this into (12). Since $\widetilde{D}^k_{it}(r)$ is orthogonal to cohort and time effects within slice k, only the treatment effect terms remain. Expanding the treatment effects across all relative times and slices:

$$\begin{aligned} \sum_k \sum_t \sum_i Y^k_{it} \widetilde{D}^k_{it}(r) = &\sum_k N^k_{g=k,t=k+r} \widetilde{D}^k_{gt}(r) \delta_{rk} \\ &+ \sum_{r' \geq 0, \neq r} \sum_k N^k_{g=k,t=k+r'} (\delta_{r'k} - \overline{\delta}'_r) \widetilde{D}^k_{k,t+r'}(r) \\ &+ \sum_k \sum_t \sum_i \widetilde{D}^k_{it}(r) \epsilon_{it} \end{aligned} \tag{14}$$

where $\overline{\delta}_{r'} = \sum_k \frac{N^k_{g=k,t=k+r'}}{N_{r'}} \delta_{r'k}$ is the sample-weighted average effect at relative time r'. The notation $\widetilde{D}^k_{gt}$ is shorthand for the common value of $\widetilde{D}$ for cohort g in time t. Thus, the $\widetilde{D}^k_{k,t+r'}(r)$ is the (generally negative) common value of $\widetilde{D}$ for cohort k in period $t + r'$.

The first term is a weighted average of δ_{rk} with weights proportional to $N^k_{g=k,t=k+r} \widetilde{D}^k_{gt}(r)$. The third term is zero in expectation from $\mathbb{E}[\epsilon_{it} \mid G_i] = 0$ for all t.

The second term is generally non-zero and is contamination coming from other treatment effects. Since the vector of relative time indicators is not interacted with the vector of slice indicators, the residualized treatment effect indicator is not orthogonal to the other treatment effect indicators (corresponding to dynamic effects in other periods $r' \neq r$) *within a slice*. It is only orthogonal unconditionally, summing over slices. Thus, while for all r', it may still co-vary with $\delta_{r'k}$ and so this term in general does not vanish.

The second term generally does not vanish because $\widetilde{D}^k_{it}(r)$ is orthogonal to other treatment indicators only on average across slices, not within each slice. That is, for any $r \neq r$, $\sum_k N^k_{g=k,t=k+r'} \widetilde{D}^k_{k,t+r'}(r) = 0$ but δ_{rk} may still covary with $\widetilde{D}^k_{gt}(r)$ across k, introducing bias in the estimated δ_r.

Each period r is contaminated by the covariance between the treatment effects $\delta_{r'k}$ and the fitted propensity scores $\widetilde{D}^k_{k,t+r'}(r)$ across slices.

Only in the special case when there is only one post-treatment period r, is there no contamination.[9] We now turn to this case.

A.3 Iterative One-Shot Aggregation

The contamination problem motivates estimating effects one relative time at a time. By sub-setting data to include only relative time r and the reference period, we eliminate contamination from other dynamic effects. We now consider iteratively aggregating dynamic treatment effects via regression.

Consider the stacked dataset, indexed by slice k as well as i, t. Then, subset the data to include *all of the control units*, but exclude treated units in relative times other than the relative time of interest r and the reference period -1.

The long regression includes slice-by-cohort and slice-by-time fixed effects and a constant. Since we only keep the relative time of interest r and the reference period -1, we do not have controls for the other dynamic effects, but those periods never appear in the regression.

Consider again the short regression of Y_{it}^k on $\widetilde{D}_{it}^k(r)$ which is the residualized indicator for relative time r.

Appealing to FWL, we can write the coefficient in the long regression using the short regression.

$$\begin{aligned}\sum_k \sum_t \sum_i Y_{it}^k \widetilde{D}_{it}^k(r) = \sum_k N_{g=k,t=k+r}^k \widetilde{D}_{gt}^k(r)\delta_{rk} \\ + \sum_k \sum_t \sum_i \widetilde{D}_{it}^k(r)\epsilon_{it}\end{aligned} \tag{15}$$

Parallel trends implies that $\mathbb{E}\left[\epsilon_{it} \mid G_i\right] = 0$ for all t, so $\mathbb{E}\left[\sum_k \sum_t \sum_i \widetilde{D}_{it}^k(r)\epsilon_{it}\right] = 0$.

Further, since we have two dimensions of fixed effects, the $\widetilde{D}_{gt}^k(r)$ have a simple form. They are time and cohort demeaned treatment indicators, with demeaning done within each slice.

$$\widetilde{D}_{gt}^k(r) = \mathbb{1}\left\{G_i = k, t = k + r\right\} - \frac{N_{G_i=k,t=k+r}^k}{N_{G_i=k}^k} - \frac{N_{G_i=k,t=k+r}^k}{N_k^k} + \frac{N_{G_i=k,t=k+r}^k}{N^k} \tag{16}$$

The dynamic effects thus have weights proportional to

$$\omega_{rk} \equiv N_{g=k,t=k+r}^k \left(1 - \frac{N_{G_i=k,t=k+r}^k}{N_{G_i=k}^k} - \frac{N_{G_i=k,t=k+r}^k}{N_k^k} + \frac{N_{G_i=k,t=k+r}^k}{N^k}\right) \tag{17}$$

These weights sum to one, but are not guaranteed to be non-negative in general.[10] Intuitively, the

[9] Of course, the contamination will be zero in expectation whenever there is no heterogeneity in treatment effects or the heterogeneity in treatment effects is "noise" in the sense that it is independent of the cohort identifier. However, in that case, heterogeneity robust estimators are not needed and the classic TWFE regression is correctly specified.

[10] To see that they sum to one, use the fact that the denominator can be simplified as follows: $\sum_k \sum_t \sum_i (\widetilde{D}_{it}^k(r))^2 =$

fixed effect structure can yield fitted values for the propensity of treatment that are greater than 1. This happens if the treated cohort k in the time of interest $k+r$ is a very large share of the slice. Weights are guaranteed to be positive if $\frac{N^k_{G_i=k,t=k+r}}{N^k_{G_i=k}} < 0.5$ and $\frac{N^k_{G_i=k,t=k+r}}{N^k_k} < 0.5$; i.e. if less than half of the treated cohort's observations are in the period of interest and less than half of the observations in the time period of interest $k+r$ are from the treated cohort.

It is possible to construct sample weights that guarantee all ω_{rk} are positive. For example, one can re-weight control cohorts and periods so that $\frac{N^k_{G_i=k,t=k+r}}{N^k_{G_i=k}} < 0.5$ and $\frac{N^k_{G_i=k,t=k+r}}{N^k_k} < 0.5$ in the reweighted sample. Note that these weights depend only on the sizes of the slice–cohort–time cells. This means they can be computed in advance, possibly after reweighting the sample. Once the ω_{rk} in Equation (17) are known, they can be rescaled by slice-specific constants to target any desired aggregation of the δ_{rg}. For instance, choosing weights proportional to $\frac{N^k_{g=k,t=k+r}}{\omega_{rk}}$ yields a cohort–size–weighted average.

$\sum_k \sum_t \sum_i D^k_{it}(r)\widetilde{D}^k_{it}(r)$ and evaluating the treatment indicator yields $\sum_k N^k_{g=k,t=k+r}\widetilde{D}^k_{gt}(r) = \sum_k \omega_{rk}$.