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

Dylan Balla-Elliott Isaac Norwich*

July 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 unnecessarily limits 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 demonstrate the approach through an application to 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.

Heterogeneity robust difference-in-differences (DiD) methods directly compare changes in treated units to changes in control units (Callaway and Sant'Anna, 2021). Estimating long-run treatment effects requires control units that remain untreated throughout the entire post-treatment window for the cohort of interest. These controls are unavailable whenever researchers (choose to) observe fewer pre-treatment periods than post-treatment periods.[1] This may arise due to data construction limits; in administrative datasets, units may only appear for a few periods before their own treatment. Or, this could be a design choice to limit the control group to only the units that are soon-to-be-treated. We show that the cohort-stacked difference-in-differences (CCDID) estimator can identify long run effects by chaining together successive soon-to-be-treated controls.[2]

When common trends and no anticipation hold across all cohorts and time periods, standard methods unnecessarily restrict the control group and recover only a subset of the parameters identified by the assumptions and data. The rank condition necessary to identify the cohort-stacked regression is weaker than the "strict balance" condition that the same control cohorts appear in the reference pre-period and the time period of interest. Intuitively, the cohort-stacked regression chains together overlapping control cohorts. 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.[3]

We demonstrate the approach through an application to earnings effects of parenthood. Following Cortés and Pan (2023), we restrict to parents within five years of first birth to ensure comparable controls. 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 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] Here we consider post-treatment as inclusive of the period in which treatment occurs.

[2] 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).

[3] Bellégo, Benatia and Dortet-Bernadet (2025) make a similar argument in unbalanced panels with attrition or rotating sampling designs. In contrast, this paper focuses specifically on staggered treatment settings where entire cohorts are observed for only a limited window before treatment. Additionally, we propose a simple "one shot" regression in contrast to their direct (multi-step) GMM estimator.

1.1 The DiD Setting

All units receive an absorbing treatment, but there is variation in treatment timing.[4] 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 corresponds 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' with 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.

1.3 Chaining Through Intermediate Periods

When no single control cohort is observed in both s and t, we can "chain" trends through intermediate periods. Suppose cohort g' exists in periods p and t, while cohort g'' exists in periods s and p, where $s < p < t$.

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

Under common trends:

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

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 to 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

Fix a treated cohort g. Then subset the data to include all observations for cohort g and only pre-treatment observations for all 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.[5] This is a testable condition. Researchers can verify whether their data structure permits chaining for desired horizons.

Within this subset, estimate:

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

[5]Formally, the design matrix has full rank if and only if all cohorts and time periods belong to the same connected set.

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 Manual aggregation through a stacked regression

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. Then the fully saturated regression is:

$$Y_{itk} = \sum_k \sum_{g'} \gamma_{g'}^k \mathbf{1}\{G_i = g'\} + \sum_k \sum_s \tau_s^k \mathbf{1}\{t = s\} + \sum_k \sum_{r \neq -1} \delta_{rk} \mathbf{1}\{G_i = k, r = t - k\} + \epsilon_{itk} \quad (8)$$

Notice now that the cohort fixed effects $\gamma_{g'}^k$, the time trends τ_s^k, 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:

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

This manual aggregation provides full control over weighting but requires additional computation. Section 2.3 demonstrates how the semi-saturated stacked regression directly estimates these aggregated parameters.

2.3 One-shot aggregation in a stacked regression

In the leading case where the target parameters are relative-time aggregates, the semi-saturated regression will recover a non-negative cohort size-weighted average of the cohort-specific dynamic effects.

The semi-saturated regression interacts the vector of slice indicators fully with the vector of time

effects and the vector of cohort effects, but not the dynamic effects of interest.

$$Y_{itk} = \sum_{k} \sum_{g'} \gamma_{g'}^{k} \mathbf{1}\{G_i = g'\} + \sum_{k} \sum_{s} \tau_{s}^{k} \mathbf{1}\{t = s\} + \sum_{r \neq -1} \delta_r \mathbf{1}\{G_i = k, r = t - k\} + \epsilon_{itk} \tag{10}$$

The δ_r coefficients are cohort size-weighted averages of the underlying cohort-specific treatment effects δ_{rg}.

$$\delta_r = \sum_{g \in \mathcal{G}} \delta_{rg} \times \mathbb{P}[G_i = g \mid G_i = k, r = t - k] \tag{11}$$

Since all other coefficients in the regression are fully interacted with the slice indicators, the relative time indicators are identified only from within-slice variation, avoiding the undesired contrasts in the canonical TWFE specification documented in the large recent literature (see Roth et al. (2023) for a review). See Appendix B for details.

The tradeoff between the one-shot and manual aggregation is one between simplicity and control. In the one-shot regression, the semi-saturated regression recovered a cohort size-weighted average of the underlying cohort-specific treatment effects δ_{rg}, which will reflect the sizes of the cohorts. In contrast, the manual aggregation of the slice-specific effects allows researchers to directly choose the weights in the aggregation, but requires the additional matrix multiplication step to compute point estimates and standard errors.

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).

We estimate long-run effects in the spirit of their Figure 1. For simplicity, we plot the effect of the first child on annual labor earnings in levels rather than percentages. We also exclude periods after 1997 to avoid gaps in the data.[6]

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 paper proposes a novel "chaining" identification strategy that expands the set of identifiable treatment effects in staggered adoption settings where conventional DiD estimators are unable to handle incomplete overlap between cohorts. By exploiting common trends assumptions across overlapping cohorts in calendar time, our approach reconstructs counterfactual trends through intermediate periods, enabling identification of treatment effects at longer horizons than existing methods allow.

The methodology demonstrates that when standard common trends and no anticipation assumptions hold—as commonly maintained in the literature—current approaches unnecessarily restrict

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

the control group and fail to fully exploit these assumptions by using only a subset of available observations. Our stacking framework jointly estimates all treatment effects for a given relative time period, while maintaining consistency with established methodologies when both approaches are feasible. The application to Cortés and Pan (2023) illustrates the practical value of this expansion, particularly in settings where researchers face natural restrictions on pre-treatment observation windows.

The key contribution lies not in imposing stronger assumptions, but in recognizing that existing assumptions, when applied consistently across all cohorts and time periods, support a broader identification strategy than previously recognized. This creates substantial practical value for empirical researchers working with administrative datasets or survey panels where pre-treatment observation windows are naturally limited or staggered.

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.

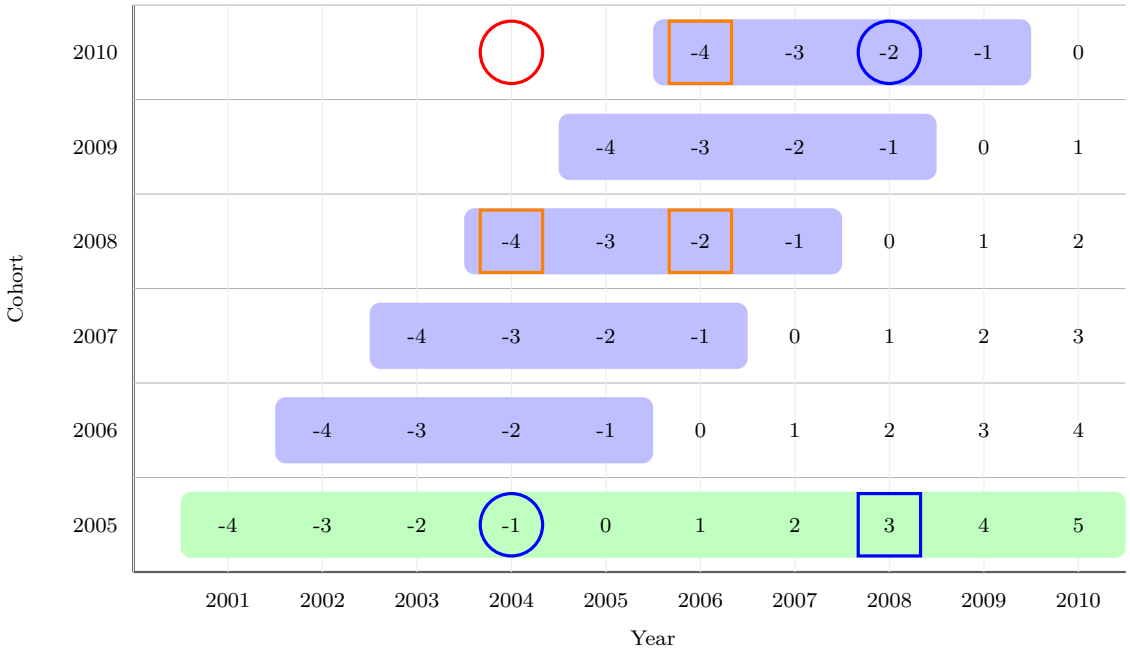
Roth, Jonathan, Pedro H.C. Sant'Anna, Alyssa Bilinski, and John Poe. 2023. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature." *Journal of Econometrics*, 235(2): 2218–2244.

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

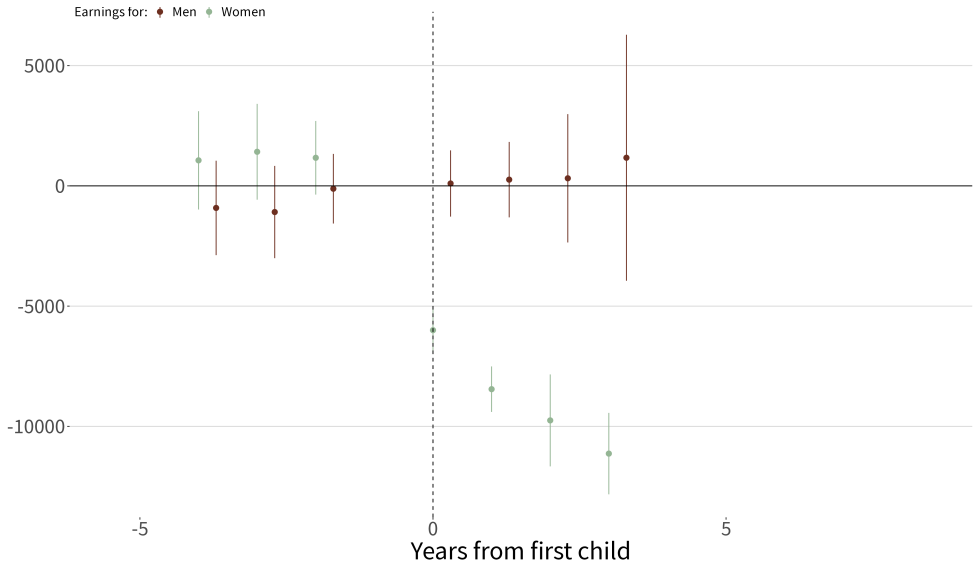
Figure 1: Example Panel Setting



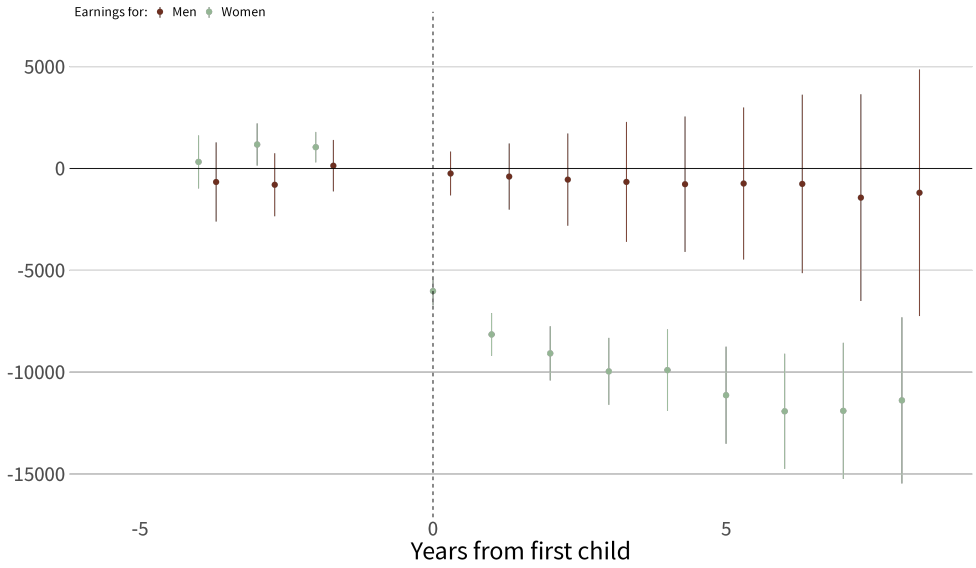
Notes: This figure shows an example panel setting with $r_{\min} = -4$ and $T_{\max} = 2010$ for cohorts $g \in \mathcal{G} = \{2005, \ldots, 2010\}$. The green shading denotes the observations for the cohort of interest, $g = 2005$, while the purple shading highlights later-treated cohorts' pre-treatment 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

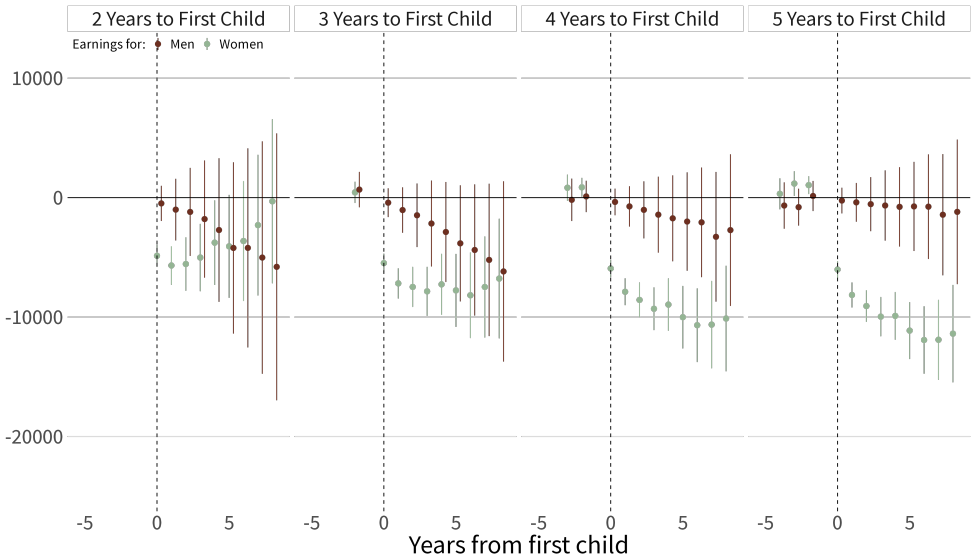


(b) Stacked DiD Estimates: Within 5 Years of First Child



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 CCID 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:

$$Y_{it} = \sum_{g'} \gamma_{g'} \mathbf{1}\{G_i = g'\} + \sum_{s} \tau_s \mathbf{1}\{t = s\} + \sum_{r \neq -1} \delta_{rg} \mathbf{1}\{G_i = g, r = t - g\} + \epsilon_{it} \tag{12}$$

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

$$\text{ATT}_r(g) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty) | G_i = g] = \delta_{rg} \tag{13}$$

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. Formally, the design matrix has full rank if and only if all cohorts and time periods belong to the same connected set.[7]

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.

[7] 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 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.

B Aggregation via semi-saturated regression

Proposition 1 (Weights in Semi-Saturated Regression)**.** *The coefficients δ_r in the semi-saturated regression*

$$Y_{itk} = \sum_k \sum_{g'} \gamma_{g'}^k \mathbf{1}\{G_i = g'\} + \sum_k \sum_s \tau_s^k \mathbf{1}\{t = s\} + \sum_{r \neq -1} \delta_r \mathbf{1}\{G_i = k, r = t - k\} + \epsilon_{itk} \tag{10}$$

are cohort size-weighted averages of the underlying cohort-specific treatment effects:

$$\delta_r = \sum_{g \in \mathcal{G}} \delta_{rg} \times \mathbb{P}[G_i = g \mid G_i = k, r = t - k] \tag{14}$$

Proof. By the Frisch-Waugh-Lovell theorem, we can partial out the slice-cohort and slice-time effects to obtain the residuals $\tilde{Y}_{itk}$. Since the remaining regressors are a vector of indicators, the coefficients on these indicators are simply difference in means, relative to the omitted group. Thus

$$\delta_r = \mathbb{E}[\tilde{Y}_{itk} \mid G_i = k, r = t - k] - \mathbb{E}[\tilde{Y}_{itk} \mid G_i \neq k \text{ or } r = -1] \tag{15}$$

Now consider the residuals conditional on (G_i, t, k):

$$\mathbb{E}[\tilde{Y}_{itk} \mid G_i = g'] \tag{16}$$

Fix a slice k' and notice that Assumption 1 (CT) and Assumption 2 (NA) imply that there are only two cases. If $t < k'$ (in the pre-period) or $g \neq k'$ (control cohorts), then $\mathbb{E}[Y_{itk} \mid G_i = g] = \gamma_g + \tau_t$ and the within-slice cohort and time indicators fit the conditional expectation function so that these residuals $\mathbb{E}[\tilde{Y}_{itk'} \mid G_i = g] = 0$. Otherwise, if $t \geq k'$ and $g = k'$ then $\mathbb{E}[Y_{itk} \mid G_i = g = k'] = \gamma_g + \tau_t + \delta_{rg}$ and after partialing out the cohort and time indicators $\mathbb{E}[\tilde{Y}_{itk'} \mid G_i = g] = \delta_{rg}$. Thus

$$\mathbb{E}[\tilde{Y}_{itk} \mid G_i = g] = \begin{cases} 0 & \text{if } t < k \text{ or } g \neq k \\ \delta_{rg} & \text{if } t \geq k \text{ and } g = k \end{cases} \tag{17}$$

Now, return to the two conditional expectations of residuals that identify δ_r and iterate expectations over (G_i, t, k). Notice that we have already conditioned on $G_i = k$ so that iterating on k is sufficient to also iterate on G_i. Then

$$\begin{aligned} \mathbb{E}[\tilde{Y}_{itk} \mid G_i = k, r = t - k] = \sum_{k'} \sum_t \mathbb{E}[\tilde{Y}_{itk'} \mid G_i = k'] \\ \times \mathbb{P}[k = k', G_i = k', t = t' \mid G_i = k, r = t - k] \end{aligned} \tag{18}$$

Notice also that since everything is conditioned on r, further conditioning on both t and k is redundant, since $\mathbb{P}[t = t' \mid r = t - k] = 0$ whenever $t' \neq r - k$. Thus we can eliminate the explicit

condition on t and simplify to

$$\begin{aligned}\mathbb{E}[\tilde{Y}_{itk} \mid G_i = k, r = t-k] = \sum_{k'} \mathbb{E}[\tilde{Y}_{i(t=r+k')k'} \mid G_i = k'] \\ \times \mathbb{P}[k = k', G_i = k' \mid G_i = k, r = t-k]\end{aligned} \tag{19}$$

Notice that $\mathbb{P}[G_i = k' \mid G_i = k] = \mathbb{P}[k = k', G_i = k' \mid G_i = k]$. Then apply Equation (17) to simplify further:

$$\mathbb{E}[\tilde{Y}_{itk} \mid G_i = k, r = t-k] = \sum_{k=g'} \delta_{rg'} \times \mathbb{P}[G_i = g' \mid G_i = k, r = t-k] \tag{20}$$

To complete the proof, notice that Equation (17) also implies that

$$\mathbb{E}[\tilde{Y}_{itk} \mid r = -1 \text{ or } G_i \neq k] = 0 \tag{21}$$

Therefore:

$$\begin{aligned}\delta_r &= \mathbb{E}[\tilde{Y}_{itk} \mid G_i = k, r = t-k] - \mathbb{E}[\tilde{Y}_{itk} \mid G_i \neq k \text{ or } r = -1] \\ &= \sum_{g \in \mathcal{G}} \delta_{rg} \times \mathbb{P}[G_i = g \mid G_i = k, r = t-k]\end{aligned} \tag{22}$$

This completes the proof that δ_r recovers a cohort size-weighted average of the underlying effects. □