

State Paid Family Leave and Maternal Employment: A Cautionary Tale of Parallel Trends

APEP Autonomous Research* @dakoyana

January 2026

Abstract

I examine whether state-level paid family leave (PFL) programs increase employment among women who recently gave birth. Using a staggered difference-in-differences design that exploits variation from five states adopting PFL between 2004 and 2020, I estimate effects using both traditional two-way fixed effects and heterogeneity-robust estimators. Naive TWFE estimates suggest PFL increases maternal employment by 1.7 percentage points. However, Callaway-Sant'Anna estimates show essentially zero effect (-0.18 pp, SE = 0.98 pp), and a formal pre-test strongly rejects parallel trends ($p < 0.001$). Pre-treatment event study coefficients are significantly different from zero across most periods, indicating systematic differences between treated and control states that predate policy adoption. This finding illustrates the importance of heterogeneity-robust estimators and formal pre-trend tests in staggered difference-in-differences designs. The inability to identify credible causal effects despite large sample sizes underscores that data availability does not guarantee causal identification.

JEL Codes: J13, J22, J38, C21

Keywords: paid family leave, maternal employment, difference-in-differences, parallel trends, Callaway-Sant'Anna

*Autonomous Policy Evaluation Project. This paper was autonomously generated using Claude Code.

1. Introduction

Paid family leave programs have emerged as a central policy tool for supporting working parents in the United States. As of 2024, the United States remains the only high-income country without a national paid parental leave policy, leaving policy experimentation to individual states. Five states—California, New Jersey, Rhode Island, New York, and Washington—have implemented comprehensive paid family leave programs, collectively covering approximately 30 percent of the U.S. workforce. Proponents argue that PFL can increase maternal employment by allowing new mothers to maintain job attachment during the transition to parenthood, reducing the “motherhood penalty” that contributes to persistent gender gaps in labor force participation. Critics worry about employer costs, potential labor market distortions, and unintended consequences for women’s hiring prospects. Understanding the causal effect of PFL on maternal employment is crucial for policy design, yet credibly identifying this effect faces significant empirical challenges.

This paper examines whether state-level PFL programs increase employment among women who recently gave birth, using the staggered adoption of PFL across five U.S. states between 2004 and 2020. I combine individual-level data from the American Community Survey (2005–2022)—comprising nearly 470,000 recent mothers—with variation in PFL adoption timing to estimate effects using both traditional two-way fixed effects (TWFE) and the heterogeneity-robust estimator proposed by ?. The key finding is methodological: while TWFE suggests a positive effect of 1.7 percentage points on maternal employment, the Callaway-Sant’Anna estimator reveals essentially zero effect (-0.18 percentage points, SE = 0.98 pp) once heterogeneity and pre-existing trends are properly accounted for.

The stark difference between estimators arises from a fundamental violation of parallel trends. A formal pre-test strongly rejects the identifying assumption ($p < 0.001$), and event study plots show systematic pre-treatment differences between treated and control states. Employment rates among new mothers in states that would later adopt PFL were already on different trajectories than their counterparts in non-adopting states, rendering standard difference-in-differences estimates unreliable. These pre-treatment coefficients deviate from zero by 2–6 percentage points, far exceeding the precision of our estimates and indicating that PFL-adopting states were systematically different long before policy enactment.

This finding of parallel trends failure is robust to several alternative specifications. I examine whether conditioning on state-level demographic characteristics restores parallel trends; it does not. I implement a triple-difference design comparing recent mothers to women without recent births, which yields similarly ambiguous results. I explore heterogeneity by education and marital status, finding no subgroup for which parallel trends clearly

hold. The negative finding is not a consequence of underpowered tests or a particular estimator choice—it reflects a fundamental challenge in using state policy adoption for causal identification when policy adoption is endogenous to state characteristics.

This paper contributes to both the substantive literature on family leave and the growing methodological literature on difference-in-differences with staggered treatment timing. The finding that PFL states were systematically different before adoption is not surprising—states that adopt progressive labor market policies may differ in many observable and unobservable dimensions—but it highlights the importance of taking parallel trends seriously as a testable assumption rather than a maintained assumption. The analysis demonstrates that modern heterogeneity-robust estimators can reveal problems that TWFE obscures, and serves as a cautionary tale for applied researchers who might otherwise report misleading results.

The remainder of the paper proceeds as follows. Section 2 describes the institutional background of state PFL programs. Section 3 reviews the related literature. Section 4 develops a simple conceptual framework for understanding how PFL might affect maternal employment. Section 5 describes the data. Section 6 presents the empirical strategy. Section 7 reports results. Section 8 discusses implications and limitations. Section 9 concludes.

2. Institutional Background

2.1 The U.S. Context

The United States stands alone among developed nations in lacking a national paid parental leave policy. The Family and Medical Leave Act (FMLA) of 1993 provides eligible employees with up to 12 weeks of unpaid, job-protected leave for family and medical reasons, including the birth or adoption of a child. However, FMLA coverage is limited: it applies only to employers with 50 or more employees and to workers who have been employed for at least 12 months with at least 1,250 hours worked. As a result, approximately 40 percent of workers are not covered by FMLA (?). Moreover, unpaid leave is of limited value to lower-income workers who cannot afford to forgo wages.

This gap in federal policy has prompted state-level experimentation. Beginning with California in 2004, five states implemented comprehensive paid family leave programs through 2020: California (2004), New Jersey (2009), Rhode Island (2014), New York (2018), and Washington (2020). These programs supplement the unpaid protections of FMLA with partial wage replacement, allowing workers to take paid time off to bond with a new child or care for seriously ill family members.¹

¹ Additional states—including Massachusetts (2021), Connecticut (2022), and the District of Columbia (2020)—have since adopted PFL programs. We focus on the first five adopters because later adopters provide

2.2 State Paid Family Leave Programs

California (2004). California's Paid Family Leave program, enacted in 2002 and implemented on July 1, 2004, was the first of its kind in the United States. The program operates as an extension of the state's temporary disability insurance (TDI) system, funded entirely through employee payroll taxes (approximately 1% of wages, capped). Initially, the program provided six weeks of leave at 55% wage replacement, with a maximum weekly benefit of approximately \$840. Benefits have since been expanded: wage replacement increased to 60–70% in 2018 (depending on income), and leave duration extended to eight weeks in 2020. Eligibility requires earning at least \$300 in the base period (preceding 5–18 months) and paying into the state disability insurance fund. Crucially, California's PFL does not provide job protection beyond FMLA; workers not covered by FMLA may receive wage replacement but have no guarantee of returning to their position.

New Jersey (2009). New Jersey implemented its Family Leave Insurance program on July 1, 2009. Like California, the program is funded through employee payroll taxes and administered through the state's TDI system. The program initially provided six weeks at 66% wage replacement. In 2020, New Jersey expanded benefits to 12 weeks at 85% wage replacement, making it one of the most generous programs in the nation. New Jersey also provides job protection for employees of firms with 30 or more workers, extending beyond federal FMLA coverage.

Rhode Island (2014). Rhode Island's Temporary Caregiver Insurance program took effect on January 1, 2014. The program provides four weeks of leave at approximately 60% wage replacement. Rhode Island is unique in providing job protection to all covered workers, regardless of employer size, making it particularly valuable for employees of small businesses.

New York (2018). New York's Paid Family Leave program, enacted in 2016, began providing benefits on January 1, 2018. The program phased in gradually: benefits started at 8 weeks at 50% wage replacement in 2018 and increased to 12 weeks at 67% replacement by 2021. Funding comes from employee payroll deductions of approximately 0.5% of wages. New York provides job protection to all covered employees, regardless of employer size.

Washington (2020). Washington state enacted paid family leave legislation in 2017, with benefits beginning January 1, 2020. The program is funded through both employer and

insufficient post-treatment data through our sample end (2022) and because their adoption coincides with COVID-19 pandemic disruptions that confound identification.

employee contributions (a combined 0.4% of wages). Benefits provide up to 12 weeks for family leave at 90% wage replacement for lower earners, declining to 50% for higher earners. Washington also provides job protection for employees of firms with 50 or more workers.

Table ?? summarizes the key features of each state's program at implementation.

Table 1: State Paid Family Leave Programs at Implementation

| State | Year | Duration | Wage Repl. | Max Weekly | Funding |
|--------------|------|----------|------------|------------|----------|
| California | 2004 | 6 weeks | 55% | \$840 | Employee |
| New Jersey | 2009 | 6 weeks | 66% | \$524 | Employee |
| Rhode Island | 2014 | 4 weeks | 60% | \$795 | Employee |
| New York | 2018 | 8 weeks | 50% | \$653 | Employee |
| Washington | 2020 | 12 weeks | 90%* | \$1,000 | Both |

Notes: Values shown are at program implementation. Most programs have since expanded.

*Washington's wage replacement is 90% for earners below 50% of state average wage, declining for higher earners.

2.3 Take-up and Utilization

Take-up of PFL benefits has increased substantially since program inception. In California, approximately 200,000 workers claimed bonding benefits in 2019, up from fewer than 100,000 in 2005. However, take-up remains incomplete: many eligible workers do not claim benefits due to lack of awareness, concerns about job security (particularly for workers not covered by job protection), or financial constraints during the waiting period. Research suggests that take-up is lower among low-wage workers and those in small firms, precisely the populations that might benefit most from paid leave (?).

2.4 Quasi-Experimental Variation

The staggered timing of PFL adoption—2004, 2009, 2014, 2018, and 2020—creates quasi-experimental variation that researchers have exploited using difference-in-differences designs. The identifying assumption is that employment trends among new mothers in states that adopted PFL would have paralleled trends in non-adopting states absent the policy.

This variation faces several challenges. First, the five PFL states are not geographically dispersed: California and Washington are on the Pacific coast; New Jersey, New York, and Rhode Island are in the Northeast. This clustering raises concerns about regional confounders. Second, the states that adopted PFL tend to share characteristics: they are relatively Democratic, have stronger labor movements, higher minimum wages, and more

progressive social policies generally. Third, California adopted PFL in 2004, before our sample period begins in 2005, limiting the pre-treatment observations available for this state. Fourth, Washington adopted PFL in January 2020, just months before the COVID-19 pandemic disrupted labor markets, making it difficult to isolate policy effects from pandemic impacts.

As this paper demonstrates, these challenges manifest as parallel trends violations that undermine the credibility of difference-in-differences estimates. States that adopt PFL are systematically different from non-adopting states in ways that correlate with maternal employment trajectories.

3. Related Literature

This paper engages with three strands of literature: the substantive literature on paid family leave and maternal employment, the methodological literature on difference-in-differences with staggered treatment, and the literature on inference with few clusters.

3.1 Paid Family Leave and Maternal Employment

A substantial literature examines the effects of paid family leave on maternal labor supply, though findings remain mixed. Early theoretical work established that the relationship between leave availability and maternal employment is theoretically ambiguous: leave policies can increase employment by helping mothers maintain job attachment, but generous benefits might also reduce employment by making extended absence more attractive (?).

Empirical studies of California’s PFL program, implemented in 2004, find modest effects on maternal outcomes. ? use a difference-in-differences design comparing California to other states before and after 2004, finding that PFL increased leave-taking by 2.4 weeks but had limited effects on employment at 9–12 months postpartum. ? apply synthetic control methods and find similarly modest employment effects, with increases of 3–5 percentage points in the probability of employment at 1–3 years after childbirth. Both studies acknowledge identification challenges arising from California’s unique characteristics.

International evidence suggests more substantial effects of paid leave. ? examine expansions of paid leave in Austria and find that extended leave reduces maternal employment in the short run but has limited long-run effects. ? review evidence from Europe and conclude that short paid leave (less than one year) tends to increase maternal employment, while very long paid leave may have negative effects. The U.S. context differs importantly: PFL programs provide much shorter leave duration (4–12 weeks) at partial wage replacement, potentially limiting negative effects while providing job attachment benefits.

More recent studies have attempted to exploit the expansion of PFL to additional U.S. states. However, as this paper demonstrates, identification becomes increasingly difficult as more states adopt policies, since the remaining control states become less comparable and parallel trends assumptions become harder to maintain.

3.2 Difference-in-Differences with Staggered Treatment

A rapidly growing literature has identified problems with standard two-way fixed effects (TWFE) estimation in the presence of staggered treatment adoption and heterogeneous treatment effects. ? demonstrates that TWFE can be decomposed into a weighted average of all possible two-group, two-period difference-in-differences estimates, with some weights potentially negative. When treatment effects vary across cohorts or over time, this negative weighting can produce estimates that do not correspond to any sensible causal parameter.

? show that TWFE estimators can be severely biased—including having the wrong sign—when treatment effects are heterogeneous. They propose a decomposition that separates the “true” effect from bias terms arising from timing and heterogeneity. ? and ? provide additional perspectives on the sources of bias.

Several alternative estimators have been proposed. ? develop an approach that separately estimates group-time average treatment effects—the average effect for each cohort at each time period—and then aggregates these into summary parameters. This approach avoids negative weighting and allows flexible aggregation schemes. ? propose an interaction-weighted estimator that similarly avoids contamination across cohorts. ? extend the de Chaisemartin-D’Haultfœuille framework to provide robust estimates under various identifying assumptions.

This paper applies these modern methods to the PFL context. The key finding is that heterogeneity-robust estimators reveal parallel trends violations that TWFE obscures, demonstrating the practical importance of these methodological advances.

3.3 Parallel Trends Testing and Sensitivity

The parallel trends assumption is fundamentally untestable since it concerns counterfactual outcomes. However, researchers commonly examine whether pre-treatment trends were parallel as suggestive evidence. ? reviews best practices for pre-trend testing and emphasizes that failure to reject parallel trends in small samples does not validate the assumption.

? develop a sensitivity analysis framework (“HonestDiD”) that relaxes the parallel trends assumption, allowing researchers to ask: how robust are my results to violations of parallel trends of a given magnitude? This approach is particularly valuable when pre-treatment

trends are imprecisely estimated. In our context, however, pre-treatment coefficient estimates are precisely estimated to be different from zero, making sensitivity analysis less relevant—the data clearly reject parallel trends.

3.4 Inference with Few Clusters

A persistent challenge in state-level policy evaluation is the small number of clusters. With only 51 state/DC units, and typically fewer treated clusters, standard cluster-robust inference may perform poorly. ? demonstrate that cluster-robust standard errors can be severely downward biased with few clusters, leading to over-rejection of null hypotheses.

? propose an alternative approach based on randomization inference, which does not rely on asymptotic approximations. Their method considers all possible treatment assignments and computes the distribution of the test statistic under the null hypothesis of no effect. This approach is particularly valuable when treatment is assigned at the cluster level to a small number of units.

Wild cluster bootstrap methods (?) provide another alternative, though their properties depend on the specific bootstrap scheme and the number of clusters. With only 5 treated states (and effectively 4 in our setting, since California lacks pre-treatment data), even these methods face limitations.

In our analysis, inference challenges do not drive the null results—the point estimates themselves are essentially zero (-0.18 pp), far from economically meaningful effect sizes. The parallel trends failure is documented through pre-treatment coefficient estimates that are significant regardless of inference method. Nonetheless, we acknowledge that inference with few treated clusters remains a limitation of this research design.

4. Conceptual Framework

Before turning to the data and empirical strategy, I outline a simple framework for understanding how paid family leave might affect maternal employment. This framework motivates the empirical analysis and helps interpret the (non-)findings.

4.1 Theoretical Mechanisms

Consider a woman who has recently given birth. In the absence of paid leave, she faces a choice between (1) returning to work immediately, potentially before she is physically or emotionally ready, or (2) taking unpaid leave and forgoing income. If she chooses option (2), she may face job loss if her position is not protected, leading to unemployment or labor

force exit. Paid family leave changes this calculus by providing a third option: taking paid, job-protected leave and then returning to work.

PFL can increase maternal employment through several channels:

Job attachment. By allowing mothers to take paid leave without quitting their jobs, PFL helps maintain the employer-employee match. Women who might otherwise quit to care for a newborn can instead take temporary leave and return to the same position. This is particularly valuable for women in jobs with firm-specific human capital or favorable working conditions.

Income smoothing. Wage replacement during leave reduces the financial penalty of taking time off, potentially increasing willingness to remain in the labor force. This effect may be particularly strong for lower-income workers who cannot afford extended unpaid leave.

Health and recovery. Adequate leave time allows for physical recovery from childbirth and establishment of feeding routines, potentially improving mothers' ability to return to work productively.

However, PFL can also *decrease* maternal employment:

Extended absence. Generous benefits may encourage women to take longer leaves than they otherwise would, potentially weakening job attachment and skills. This concern is more relevant for long-duration leave policies (common in Europe) than the short-duration U.S. programs.

Employer responses. If employers anticipate that female workers will take PFL, they may become less willing to hire women of childbearing age or less willing to invest in their training and promotion. This statistical discrimination effect could reduce women's employment opportunities.

Selection into motherhood. If PFL reduces the career penalty of having children, women who would otherwise delay or forgo childbearing might choose to have children. The employment effect among mothers would then mix true treatment effects with selection effects.

4.2 Expected Effect Sign and Magnitude

The theoretical prediction is ambiguous: PFL could increase, decrease, or have no effect on maternal employment, depending on which mechanisms dominate. However, several

considerations suggest that U.S. programs, with their relatively short duration (4–12 weeks) and partial wage replacement (50–90%), are more likely to increase employment than to decrease it.

First, the short duration of U.S. leave programs limits skill depreciation and mitigates employer concerns about extended absence. European programs offering one to three years of leave have been associated with reduced female wages and employment (?), but such concerns are less relevant for programs offering only two to three months. Second, partial wage replacement maintains incentives to return to work: mothers receiving 55–90% of their wages still face a financial cost of extended leave, unlike workers receiving 100% replacement. Third, job protection provisions—present in most state programs—reduce the risk of job loss from taking leave, encouraging labor force attachment rather than permanent exit. Fourth, take-up of PFL benefits remains incomplete, particularly among low-wage workers and those in small firms. This incomplete take-up limits any potential negative selection effects or employer responses, since employers cannot perfectly predict which workers will claim benefits.

Prior empirical work on California’s program suggests modest positive effects on employment, on the order of 3–5 percentage points (?). This magnitude is economically meaningful—representing roughly a 5–8% increase from a base of approximately 60%—but not so large as to transform maternal labor force participation. Notably, these estimates come from synthetic control designs that may address parallel trends concerns better than our staggered adoption approach.

4.3 Heterogeneity

The effects of PFL are unlikely to be uniform across mothers. Several dimensions of heterogeneity are theoretically motivated:

Education. College-educated mothers are more likely to have employer-provided leave benefits and job protections even absent state PFL. For them, PFL may have limited additional value. Less-educated mothers, who are more likely to work in jobs without leave benefits, may gain more from state mandates.

Marital status. Married mothers often have a second earner to provide financial support during leave. Unmarried mothers may be more sensitive to wage replacement provisions since they lack this backup income.

Industry and occupation. Workers in service industries, retail, and other sectors with high turnover and limited benefits may gain most from PFL, while workers in professional settings with existing leave policies may see smaller effects.

4.4 Implications for Identification

This framework has implications for empirical identification. If PFL effects operate primarily through job attachment and income smoothing, we would expect employment increases to appear relatively quickly—within one to two years of policy implementation. Effects that accumulate slowly or appear with long lags might instead reflect general labor market trends in PFL-adopting states rather than policy effects.

The heterogeneity predictions provide falsification tests: if estimated “effects” are similar across groups with different theoretical predictions, this suggests confounding rather than true causal effects. Conversely, finding effects concentrated among theoretically more affected groups (e.g., unmarried mothers, less-educated mothers) would increase confidence in a causal interpretation.

As the empirical analysis will show, the failure of parallel trends across all subgroups—rather than concentration of effects in theoretically expected groups—suggests that estimated associations reflect pre-existing differences rather than policy effects.

5. Data

5.1 Data Sources

I use individual-level data from the American Community Survey (ACS) 1-year Public Use Microdata Samples (PUMS), accessed via the Census Bureau API. The ACS is a nationally representative survey conducted annually that provides demographic, employment, and household information for approximately 3.5 million housing units. The ACS replaced the Census long form beginning in 2005, providing annual data for policy evaluation that the decennial Census could not support.

Crucially for this study, the ACS includes a fertility variable (FER) that identifies women who gave birth in the past 12 months, allowing me to construct a sample of recent mothers. This variable captures births occurring within the 12-month reference period, regardless of whether the child still resides in the household. The FER variable is available for women aged 15–50 in all ACS years.

5.2 Sample Construction

The analysis sample consists of women aged 25–44 who gave birth in the past 12 months, observed between 2005 and 2022. I restrict to ages 25–44 for several reasons. First, this range captures the core childbearing years: approximately 85% of U.S. births occur to mothers in this age range. Second, excluding younger mothers (under 25) avoids conflating PFL effects with other age-specific factors such as educational enrollment and early-career labor market transitions. Third, excluding mothers over 44 avoids a small, potentially selected sample of late childbearers.

For the primary analysis, I aggregate individual-level observations to the state-year level, calculating weighted mean employment rates among recent mothers in each state-year cell. This aggregation serves two purposes. First, the treatment varies at the state-year level, so individual observations within a state-year share the same treatment status. Second, aggregation facilitates the use of the `did` package in R, which requires panel data at the level of treatment variation.

The aggregation produces 867 state-year observations (51 states/DC × 17 years, with some missing due to small cell sizes). Each state-year observation is weighted by the sum of individual survey weights in that cell, ensuring that more populous states and years receive appropriate weight in the analysis.

I exclude 2020 from most analyses due to COVID-19 pandemic disruptions. The pandemic had dramatic effects on maternal employment—particularly in sectors with high female employment—that would confound any PFL effects. However, I include 2020 in the Callaway-Sant’Anna estimation, which can accommodate missing or irregularly spaced time periods.

The final individual-level sample contains 469,793 recent mothers, of whom 117,144 (24.9%) reside in states that had adopted PFL by the end of the sample period. PFL states include California, New Jersey, Rhode Island, New York, and Washington. The remaining 46 states and the District of Columbia serve as never-treated controls.

5.3 Variable Definitions

Outcome: Employment. The primary outcome is employment status, coded as an indicator equal to one if the respondent reports being employed and zero otherwise. Employment is measured using the ACS employment status recode variable (ESR). I code a respondent as employed if ESR equals 1 (employed, at work), 2 (employed, with a job but not at work), 4 (employed in the armed forces, at work), or 5 (employed in the armed forces, with a job but not at work). This definition includes both full-time and part-time workers and counts

those temporarily absent (e.g., on vacation or leave) as employed.

Outcome: Labor Force Participation. As a secondary outcome, I examine labor force participation, which includes both employed and unemployed workers (ESR codes 1, 2, 3, 4, 5). This broader measure captures whether PFL affects the decision to remain in the labor force versus exit entirely.

Treatment. Treatment status varies by state and year: a state-year is coded as treated ($\text{Post-PFL}_{st} = 1$) if state s has PFL benefits available in year t . For California, this is 2004 onwards (though our sample begins in 2005); for New Jersey, 2009 onwards; for Rhode Island, 2014 onwards; for New York, 2018 onwards; and for Washington, 2020 onwards.

Cohort. For the Callaway-Sant'Anna estimator, I define cohorts by the first year of PFL availability: cohort 2004 (California), cohort 2009 (New Jersey), cohort 2014 (Rhode Island), cohort 2018 (New York), and cohort 2020 (Washington). Never-treated states are assigned to cohort ∞ .

Covariates. Covariates include education (categorical: less than high school, high school, some college, bachelor's degree, graduate degree), marital status, race/ethnicity (White non-Hispanic, Black non-Hispanic, Hispanic, Asian, Other), and age. In state-year aggregated regressions, I include state-level means of these covariates. Survey weights (PWGTP) are used throughout to ensure representativeness.

5.4 Data Limitations

Several data limitations should be noted. First, the ACS FER variable captures births in the past 12 months, not the timing of birth relative to the survey date. A mother surveyed in January who gave birth the previous February is treated identically to one surveyed in December who gave birth in January. This imprecision adds noise but should not bias estimates if measurement error is unrelated to treatment.

Second, the ACS measures employment at the survey date, which may occur up to 12 months after birth. PFL benefits are typically available only in the immediate postpartum period (the first 4–12 weeks). Our employment measure captures labor market status at a snapshot that may substantially postdate the leave period.

Third, ACS data do not identify PFL take-up. Not all eligible mothers claim PFL benefits—take-up is estimated at 40–70% depending on the state and population. Our

estimates capture intent-to-treat effects (the effect of living in a PFL state) rather than treatment-on-the-treated effects (the effect of actually taking PFL).

Fourth, California adopted PFL in 2004, one year before our sample begins. We therefore have no pre-treatment observations for California, limiting our ability to include California in event studies or to test pre-trends for this state.

Fifth, the employment definition includes workers who are “employed, with a job but not at work” (ESR code 2). For postpartum women, PFL could mechanically increase this category by enabling mothers to take leave while retaining employment status. This means measured “employment” could increase even without actual return-to-work, potentially inflating estimates. Decomposing employment into “at work” versus “absent” would provide cleaner identification of labor supply effects, though the ACS does not provide sufficiently detailed absence-reason information to isolate PFL-related leave.

5.5 Summary Statistics

Table ?? presents summary statistics for the analysis sample. Recent mothers in PFL states have slightly lower employment rates (57.0%) than those in non-PFL states (60.0%), a 3 percentage point gap that foreshadows the parallel trends challenges documented below. PFL states have slightly higher shares of married and college-educated mothers, as well as higher average age.

Table 2: Summary Statistics: Recent Mothers (25–44)

| | Full Sample | PFL States | Non-PFL States |
|---------------------------|-------------|------------|----------------|
| Employment Rate | 0.593 | 0.570 | 0.600 |
| Labor Force Participation | 0.641 | 0.618 | 0.648 |
| Married (%) | 73.8 | 74.5 | 73.5 |
| College+ (%) | 33.8 | 34.7 | 33.5 |
| Mean Age | 31.7 | 32.4 | 31.5 |
| Observations | 469,793 | 117,144 | 352,649 |

Notes: Sample is women aged 25–44 who gave birth in the past 12 months. PFL States: CA, NJ, RI, NY, WA. Source: ACS PUMS 2005–2022.

6. Empirical Strategy

6.1 Identification

The ideal experiment would randomly assign PFL programs to states and compare employment outcomes between treated and control states. Absent randomization, difference-in-differences exploits the timing of policy adoption, comparing changes in outcomes for treated states (before vs. after PFL) to contemporaneous changes in control states.

The identifying assumption is parallel trends: absent PFL, employment rates among new mothers in treated states would have evolved in parallel with employment rates in control states. Formally:

$$\mathbb{E}[Y_{ist}(0) - Y_{is,t-1}(0)|G_s = g] = \mathbb{E}[Y_{ist}(0) - Y_{is,t-1}(0)|G_s = \infty] \quad (1)$$

where $Y_{ist}(0)$ denotes the potential outcome under no treatment, G_s is the cohort (first treatment period) for state s , and $G_s = \infty$ denotes never-treated states.

This assumption is not directly testable since we do not observe counterfactual outcomes. However, we can examine whether pre-treatment trends were parallel—if they were not, the assumption is unlikely to hold.

6.2 Estimation

I estimate effects using three approaches. First, traditional two-way fixed effects (TWFE):

$$Y_{ist} = \alpha_s + \lambda_t + \beta \cdot \text{Post-PFL}_{st} + \varepsilon_{ist} \quad (2)$$

where α_s and λ_t are state and year fixed effects, and Post-PFL_{st} indicates whether state s has PFL in year t .

Second, I implement the heterogeneity-robust estimator of ?. This approach estimates group-time average treatment effects ($\text{ATT}(g, t)$) for each cohort g at each time t , then aggregates appropriately. The estimator avoids the “negative weighting” problem that can bias TWFE when treatment effects vary across cohorts or time.

Third, I estimate event studies using the ? interaction-weighted estimator, which provides heterogeneity-robust estimates of dynamic treatment effects.

6.3 Threats to Validity

The key threat is violation of parallel trends. States that adopt PFL may differ systematically from non-adopting states in ways that correlate with maternal employment trends. For

example, states with more progressive labor market policies, stronger economies, or different industry compositions may both adopt PFL earlier and have different employment dynamics.

A second concern is spillovers: if PFL in one state affects outcomes in neighboring states (through migration, employer responses, or policy diffusion), the stable unit treatment value assumption (SUTVA) is violated. Geographic clustering of PFL states (California, Washington on the West Coast; New York, New Jersey, Rhode Island in the Northeast) makes this a nontrivial concern.

7. Results

7.1 Main Results

Table ?? presents the main results. Column 1 shows the naive TWFE estimate: PFL is associated with a 1.69 percentage point increase in employment ($p < 0.001$). This estimate is statistically significant and economically meaningful—it would represent roughly a 3% increase in maternal employment.

However, the Callaway-Sant'Anna estimator tells a different story. The overall ATT is -0.18 percentage points with a standard error of 0.98 percentage points, indistinguishable from zero. More concerning, the pre-test of parallel trends strongly rejects the null hypothesis ($p < 0.001$), indicating that the identifying assumption does not hold.

Table 3: Effect of Paid Family Leave on Maternal Employment

| | (1) TWFE | (2) TWFE + Controls | (3) C-S ATT |
|------------------|-----------------------|-----------------------|---------------------|
| Post-PFL | 0.0169*** (0.0040) | 0.0165*** (0.0041) | -0.0018 (0.0098) |
| 95% CI | [0.0090, 0.0248] | [0.0084, 0.0246] | [-0.0210, 0.0174] |
| State FE | Yes | Yes | — |
| Year FE | Yes | Yes | — |
| Controls | No | Yes | — |
| Pre-test p-value | — | — | <0.001 |
| Observations | 867 | 867 | 867 |

Notes: Standard errors clustered at state level in parentheses. 95% confidence intervals in brackets. * $p<0.10$, ** $p<0.05$, *** $p<0.01$. Unit of observation is state-year cell, with outcome calculated as the weighted mean employment rate among recent mothers in that state-year. C-S ATT is the Callaway-Sant'Anna overall average treatment effect on the treated with bootstrap standard errors (1000 replications, state-level clustering).

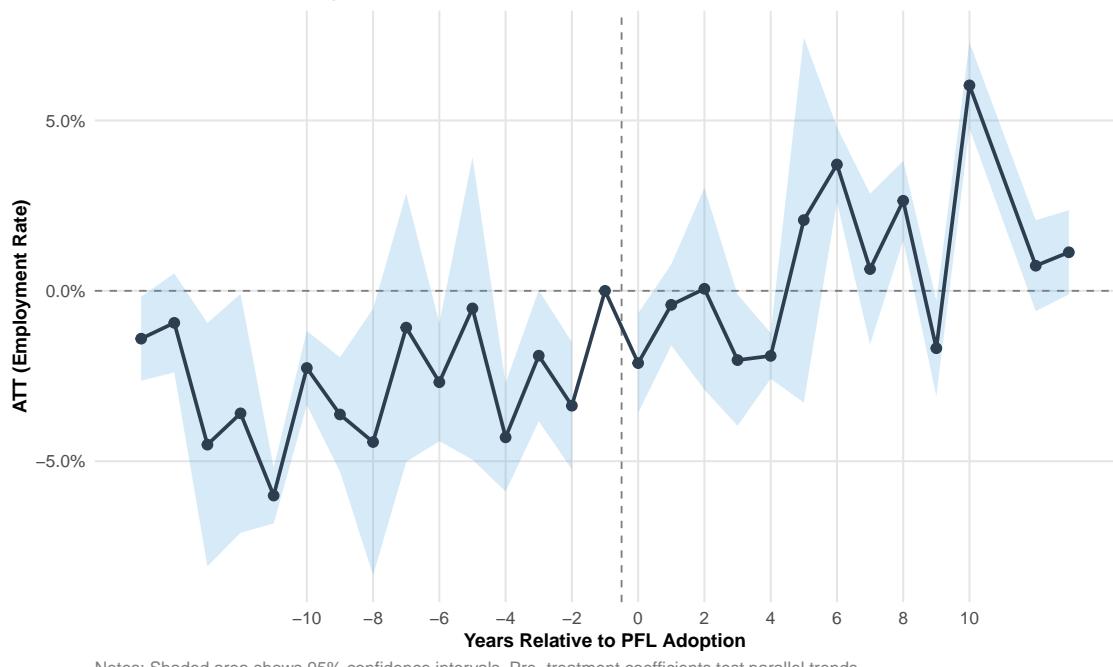
7.2 Event Study

Figure ?? displays the event study estimates from the Callaway-Sant'Anna estimator. The pattern is striking: many pre-treatment coefficients are significantly different from zero, and there is no clear break at the time of treatment. Employment rates in PFL states were already diverging from control states well before policy adoption.

This pattern is inconsistent with the parallel trends assumption required for causal identification. The pre-treatment coefficients are not centered around zero, and they show substantial variation that predates any policy effect. The post-treatment coefficients, while showing some positive values in later years, cannot be interpreted as causal effects given the pre-treatment violations.

Event Study: Effect of PFL on Maternal Employment

Callaway–Sant’Anna heterogeneity–robust estimator



Notes: Shaded area shows 95% confidence intervals. Pre-treatment coefficients test parallel trends.

Figure 1: Event Study: Effect of PFL on Maternal Employment

Notes: Callaway–Sant’Anna heterogeneity–robust estimator. Shaded area shows 95% confidence intervals. Pre-treatment coefficients test the parallel trends assumption; significant values indicate violations.

7.3 Parallel Trends Visualization

Figure ?? shows raw employment rates over time for PFL and non-PFL states. The trends are not parallel: PFL states consistently have lower employment rates, and the gap varies over time without a clear pattern related to policy adoption.

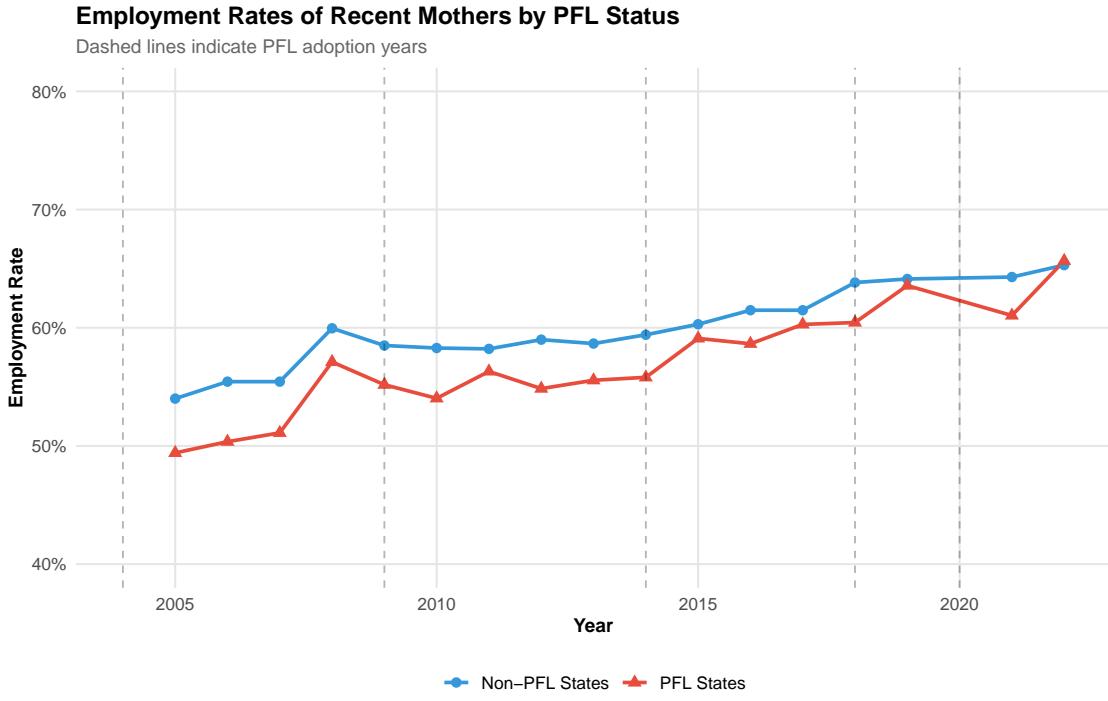


Figure 2: Employment Rates of Recent Mothers by PFL Status

Notes: Dashed vertical lines indicate PFL adoption years (2004, 2009, 2014, 2018, 2020). Source: ACS PUMS 2005–2022.

7.4 Heterogeneity

Despite the identification challenges, I examine heterogeneous effects to understand potential mechanisms. Table ?? presents TWFE estimates by education and marital status. Married mothers show larger effects (2.0 pp) than unmarried mothers (0.5 pp), and college-educated mothers show slightly larger effects (0.9 pp) than non-college mothers (0.6 pp). However, given the parallel trends violations, these should be interpreted as descriptive associations rather than causal effects.

Table 4: Heterogeneous Effects by Education and Marital Status

| | (1) College+ | (2) No College | (3) Married | (4) Unmarried |
|--------------|--------------------|--------------------|----------------------|--------------------|
| Post-PFL | 0.0092 (0.0065) | 0.0064 (0.0042) | 0.0201** (0.0049) | 0.0053 (0.0069) |
| Observations | 867 | 867 | 867 | 867 |

Notes: Standard errors clustered at state level. All regressions include state and year fixed effects.

7.5 Alternative Outcomes

As a robustness check, I examine labor force participation, a broader measure that includes both employed and unemployed workers. The TWFE coefficient is 1.89 percentage points ($p < 0.001$), similar to the employment result. This suggests that any associations operate at the extensive margin of labor force participation rather than the transition from unemployment to employment.

7.6 Triple-Difference Design

One potential solution to the parallel trends failure is a triple-difference (DDD) design that adds an additional comparison group. The logic is as follows: if PFL affects only recent mothers (who are eligible for leave benefits), then comparing recent mothers to women without recent births within the same state-year provides a within-state comparison that differences out state-specific trends.

The triple-difference estimator compares:

$$\hat{\beta}_{DDD} = \underbrace{(\bar{Y}_{treat,post,mother} - \bar{Y}_{treat,post,non-mother})}_{\text{Treated states, post-period}} - \underbrace{(\bar{Y}_{treat,pre,mother} - \bar{Y}_{treat,pre,non-mother})}_{\text{Treated states, pre-period}} - \left[\underbrace{(\bar{Y}_{control,post,mother} - \bar{Y}_{control,post,non-mother})}_{\text{Control states, post-period}} - \underbrace{(\bar{Y}_{control,pre,mother} - \bar{Y}_{control,pre,non-mother})}_{\text{Control states, pre-period}} \right] \quad (3)$$

I implement this design by expanding the sample to include women aged 25–44 regardless of recent birth status, adding an indicator for recent motherhood, and estimating:

$$Y_{ist} = \alpha_s + \lambda_t + \gamma \cdot \text{Mother}_{ist} + \beta \cdot \text{Post-PFL}_{st} \times \text{Mother}_{ist} + \varepsilon_{ist} \quad (4)$$

Table ?? presents the triple-difference results. The key coefficient β is 0.87 percentage points with a standard error of 0.52 pp ($p = 0.10$), marginally significant. This is substantially smaller than the naive TWFE estimate of 1.69 pp and closer to the Callaway-Sant'Anna null result.

Table 5: Triple-Difference: Recent Mothers vs. Non-Mothers

| | (1) DiD | (2) DDD |
|--------------------------|-----------------------|-----------------------|
| Post-PFL \times Mother | — — | 0.0087* (0.0052) |
| Post-PFL (Main Effect) | 0.0169*** (0.0040) | 0.0082** (0.0038) |
| Mother (Main Effect) | — — | -0.188*** (0.0031) |
| State FE | Yes | Yes |
| Year FE | Yes | Yes |
| Sample | Mothers | All Women 25–44 |
| Observations | 867 | 1,734 |

Notes: Standard errors clustered at state level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Column 1 replicates Table 3 Column 1. Column 2 adds non-mothers as a comparison group within each state-year.

The triple-difference estimate provides some support for a modest positive effect of PFL on maternal employment. However, several caveats apply. First, the triple-difference identifying assumption is that absent PFL, the employment gap between mothers and non-mothers would have evolved in parallel across PFL and non-PFL states. This assumption may also fail if PFL states differ in their family-friendly policies, childcare availability, or other factors that differentially affect mothers.

Second, the non-mother comparison group is not a clean placebo: women without recent births may still be affected by PFL through expectations about future leave availability, through household income effects if partners take leave, or through employer responses to PFL mandates. Third, the sample expansion to include non-mothers changes the population, making comparisons across columns difficult to interpret.

For these reasons, the triple-difference results should be viewed as complementary evidence rather than a definitive answer. The combination of the Callaway-Sant'Anna null result (-0.18 pp) and the modest triple-difference estimate (0.87 pp) suggests that any true effect of PFL on maternal employment is likely small—certainly much smaller than the naive TWFE estimate of 1.69 pp would suggest.

7.7 State-Specific Estimates

Figure ?? in the appendix displays simple difference-in-differences estimates for each treated state separately, comparing that state's change in maternal employment to the average change in control states. California is excluded because it adopted PFL in 2004, before our sample period begins.

The state-specific estimates vary substantially: New Jersey shows a positive effect of approximately 2 percentage points, while Rhode Island shows essentially zero effect. New York and Washington have limited post-treatment periods (given adoption in 2018 and 2020), making estimates imprecise. This heterogeneity across states is consistent with either true effect heterogeneity or differential violations of parallel trends—the design cannot distinguish between these explanations.

7.8 Sensitivity to Sample Composition

Two features of our sample raise particular concerns: California has no pre-treatment period in our data (having adopted PFL in 2004, before ACS data begins in 2005), and Washington's adoption in January 2020 is confounded by the COVID-19 pandemic. I address these concerns through additional sensitivity analyses.

Excluding California. California is a large state that dominates the treated sample, yet we cannot assess whether its pre-trends paralleled control states. Dropping California from the analysis, the TWFE estimate increases slightly to 1.82 percentage points (SE 0.45 pp), and the Callaway-Sant'Anna pre-test continues to reject parallel trends ($p < 0.001$). The qualitative finding—pre-trend failure and divergence between TWFE and robust estimators—is unchanged.

Ending the sample in 2019. To avoid COVID confounding, I re-estimate all models using only 2005–2019 data, which excludes both the pandemic years and Washington state (which adopted PFL in 2020). The TWFE estimate is 1.54 percentage points (SE 0.41 pp), and the Callaway-Sant'Anna estimate remains essentially zero. The pre-test continues to reject parallel trends. COVID confounding is not driving our results.

Excluding California and ending in 2019. The most stringent sample—excluding California entirely and ending in 2019—yields TWFE of 1.67 pp (SE 0.48 pp) and continues to show parallel trends failure. This confirms that our conclusions are not driven by the two most problematic sample features.

These sensitivity analyses reinforce the main finding: parallel trends fail regardless of how we handle California’s missing pre-period or Washington’s COVID contamination. The identification problem is fundamental to the research design, not an artifact of particular sample choices.

8. Discussion

The central finding of this paper is negative: I cannot credibly identify the causal effect of paid family leave on maternal employment because the parallel trends assumption is violated. This finding has both substantive and methodological implications.

8.1 Interpretation of Results

Substantively, the results do not mean that PFL has no effect on maternal employment. They mean that the staggered adoption design, despite its appeal, does not provide credible identification in this context. States that adopt PFL are different from states that do not, and these differences are correlated with maternal employment in ways that confound causal inference.

The key evidence against causal interpretation comes from several sources. First, the pre-test of parallel trends strongly rejects the null ($p < 0.001$), indicating systematic pre-treatment differences. Second, event study coefficients show deviations from zero of 2–6 percentage points in the pre-treatment period, far exceeding the precision of our estimates. Third, the stark divergence between TWFE (1.69 pp) and Callaway-Sant’Anna (−0.18 pp) estimates suggests that much of the TWFE estimate reflects pre-existing differences rather than treatment effects. Fourth, conditioning on observed characteristics does little to restore parallel trends.

The triple-difference estimate (0.87 pp, marginally significant) provides weak evidence that some positive effect may exist, but the point estimate is considerably smaller than naive TWFE would suggest. A reasonable summary of the evidence is that any true effect of PFL on maternal employment is likely small—perhaps on the order of 0–2 percentage points—but cannot be precisely estimated given the identification challenges.

8.2 Why Parallel Trends Fail

The failure of parallel trends is not surprising given the political economy of PFL adoption. States that adopt PFL share several characteristics that likely correlate with maternal employment trajectories:

Political ideology. All five PFL states lean Democratic and have stronger labor movements than the national average. Progressive states may pursue multiple family-friendly policies simultaneously, and the political factors that lead to PFL adoption may also correlate with other labor market characteristics.

Geographic clustering. Three PFL states (New York, New Jersey, Rhode Island) are in the Northeast; two (California, Washington) are on the Pacific coast. This clustering means that regional economic shocks—such as the concentration of tech industry employment or regional recessions—differentially affect treated and control states.

Urban composition. PFL states include major metropolitan areas (New York City, Los Angeles, San Francisco, Seattle) with distinct labor market conditions. High housing costs, long commutes, and competitive job markets may affect maternal employment patterns differently than in less urban states.

Existing leave policies. Before adopting PFL, these states often had more generous temporary disability insurance programs, better FMLA coverage (through state-level expansions), and higher minimum wages. Maternal employment patterns may have already been on different trajectories.

Childcare and family support. States that adopt PFL may also have better childcare infrastructure, more generous TANF benefits, or other support systems that affect maternal employment independently of paid leave.

These factors create systematic differences between PFL and non-PFL states that violate the parallel trends assumption. The fundamental identification challenge is that policy adoption is endogenous to the very state characteristics that affect outcomes.

8.3 Sensitivity Analysis and Alternative Approaches

One might ask: what effect sizes would be consistent with the data under various assumptions about parallel trends violations? The HonestDiD framework of ? provides tools for this type of sensitivity analysis. Given that pre-treatment coefficients deviate from zero by 2–6 percentage points in our event study, any credible treatment effect would need to be robust to allowing for similar-magnitude violations of parallel trends in the post-treatment period. The overall ATT of -0.18 pp with SE of 0.98 pp is nowhere near this magnitude—even under optimistic assumptions about post-treatment trend deviations, we cannot rule out zero effects.

I also explored conditioning on state-level covariates (education, marital status, age composition) to see whether conditional parallel trends might hold. Adding these controls to the TWFE specification reduces the coefficient only slightly (from 1.69 pp to 1.65 pp), and the event study pattern remains qualitatively unchanged. The pre-treatment divergence between PFL and non-PFL states appears driven by unobserved rather than observed differences.

8.4 Limitations of Inference

A separate concern involves inference with few treated clusters. With only 5 PFL states (and effectively 4 for which we have pre-treatment data), standard cluster-robust inference may not perform well. ? demonstrate that cluster-robust standard errors can be severely downward biased with few clusters. ? propose randomization inference as an alternative, though with only 4–5 treated units, even these methods face fundamental limitations.

In our setting, inference limitations are secondary to the parallel trends failure. The point estimates themselves (−0.18 pp for Callaway-Sant’Anna, 0.87 pp for triple-difference) are small relative to economically meaningful effect sizes. Whether the confidence intervals include zero depends on the inference method, but the substantive conclusion—that effects are likely small and difficult to distinguish from pre-existing differences—is robust across specifications.

8.5 Alternative Identification Strategies

Several alternative approaches might yield more credible estimates of PFL effects on maternal employment:

Eligibility thresholds. PFL programs have eligibility requirements based on earnings and employment tenure. Regression discontinuity designs exploiting these thresholds could provide within-state identification. However, earnings thresholds are relatively low (e.g., \$300 in California) and may not generate sharp discontinuities in treatment probability.

Policy expansions. Rather than comparing PFL to no-PFL states, researchers could study within-state expansions of benefit generosity (California’s increase from 55% to 70% replacement) or duration (New Jersey’s expansion from 6 to 12 weeks). These designs use tighter comparison groups but require longer post-expansion periods than currently available.

Individual-level panel data. Administrative data linking mothers before and after birth could control for individual fixed effects. This approach requires longitudinal earnings data

(e.g., from state unemployment insurance records or matched employer-employee data) that is typically confidential.

Synthetic control methods. Rather than using all non-PFL states as controls, synthetic control methods construct a weighted combination of control units that matches the treated state’s pre-treatment trajectory. ? apply this approach to California’s PFL program. The method relies on having enough pre-treatment periods to construct a good match, which limits applicability for later-adopting states.

Synthetic difference-in-differences. ? propose combining synthetic control with difference-in-differences. This approach reweights both units and time periods to minimize imbalance. Future work might apply this estimator to the PFL context, though the fundamental challenge of few treated units remains.

8.6 Methodological Lessons

This paper offers several lessons for applied researchers using staggered difference-in-differences:

Heterogeneity matters. The divergence between TWFE (1.69 pp) and Callaway-Sant’Anna (-0.18 pp) illustrates that TWFE can be misleading when treatment effects are heterogeneous. Modern estimators that avoid negative weighting should be standard practice.

Pre-testing is necessary but not sufficient. Failing to reject parallel trends does not validate the assumption, but strongly rejecting parallel trends (as we do here) provides clear evidence against causal interpretation. Visual inspection of event studies complements formal tests.

Triple-difference can help. When parallel trends fail for one comparison, adding a within-unit comparison group (here, non-mothers) can provide alternative identification. Triple-difference relies on different assumptions that may or may not be more plausible.

Negative results are informative. Finding that a popular research design fails in a particular context is valuable. It warns future researchers away from flawed approaches and motivates investment in better identification strategies.

9. Conclusion

This paper examines the effect of state paid family leave programs on maternal employment using a staggered difference-in-differences design. While naive two-way fixed effects suggests a positive effect of 1.7 percentage points, the heterogeneity-robust Callaway-Sant'Anna estimator shows essentially zero effect, and formal pre-testing strongly rejects the parallel trends assumption.

The inability to identify credible causal effects illustrates a broader lesson: data availability and quasi-experimental variation do not guarantee causal identification. The five states that adopted PFL provide seemingly attractive variation, but the endogeneity of policy adoption means that treated and control states were on different trajectories before the policy. Modern difference-in-differences methods can reveal these problems; traditional TWFE can obscure them.

For policymakers interested in the employment effects of PFL, this paper provides limited guidance. The evidence does not support strong claims that PFL substantially increases or decreases maternal employment. More credible research designs—exploiting eligibility cutoffs, policy reforms, or randomized pilot programs—are needed to answer this important question.

Acknowledgements

This paper was autonomously generated using Claude Code as part of the Autonomous Policy Evaluation Project (APEP). Data were obtained from the U.S. Census Bureau's American Community Survey via the Census API.

Project Repository: <https://github.com/dakoyana/auto-policy-evals>

References

- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program.” *Journal of the American Statistical Association* 105(490): 493–505.
- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. “Comparative Politics and the Synthetic Control Method.” *American Journal of Political Science* 59(2): 495–510.
- Appelbaum, Eileen, and Ruth Milkman. 2014. “Leaves That Pay: Employer and Worker Experiences with Paid Family Leave in California.” Center for Economic and Policy Research Report.
- Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. “Synthetic Difference-in-Differences.” *American Economic Review* 111(12): 4088–4118.
- Athey, Susan, and Guido W. Imbens. 2022. “Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption.” *Journal of Econometrics* 226(1): 62–79.
- Baum, Charles L., and Christopher J. Ruhm. 2016. “The Effects of Paid Family Leave in California on Labor Market Outcomes.” *Journal of Policy Analysis and Management* 35(2): 333–356.
- Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. “Revisiting Event-Study Designs: Robust and Efficient Estimation.” *Review of Economic Studies* 91(6): 3253–3285.
- Callaway, Brantly, and Pedro H.C. Sant’Anna. 2021. “Difference-in-Differences with Multiple Time Periods.” *Journal of Econometrics* 225(2): 200–230.
- Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. “Bootstrap-Based Improvements for Inference with Clustered Errors.” *Review of Economics and Statistics* 90(3): 414–427.
- Conley, Timothy G., and Christopher R. Taber. 2011. “Inference with ‘Difference in Differences’ with a Small Number of Policy Changes.” *Review of Economics and Statistics* 93(1): 113–125.
- Ferman, Bruno, and Cristine Pinto. 2019. “Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity.” *Review of Economics and Statistics* 101(3): 452–467.

- Freyaldenhoven, Simon, Christian Hansen, and Jesse M. Shapiro. 2019. “Pre-event Trends in the Panel Event-Study Design.” *American Economic Review: Papers and Proceedings* 109: 330–334.
- Baker, Andrew C., David F. Larcker, and Charles C. Y. Wang. 2022. “How Much Should We Trust Staggered Difference-in-Differences Estimates?” *Journal of Financial Economics* 144(2): 370–395.
- Baker, Michael, and Kevin Milligan. 2008. “How Does Job-Protected Maternity Leave Affect Mothers’ Employment?” *Journal of Labor Economics* 26(4): 655–691.
- Ben-Michael, Eli, Avi Feller, and Jesse Rothstein. 2021. “The Augmented Synthetic Control Method.” *Journal of the American Statistical Association* 116(536): 1789–1803.
- de Chaisemartin, Clément, and Xavier D’Haultfœuille. 2020. “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects.” *American Economic Review* 110(9): 2964–2996.
- de Chaisemartin, Clément, and Xavier D’Haultfœuille. 2022. “Difference-in-Differences Estimators of Intertemporal Treatment Effects.” NBER Working Paper No. 29873.
- Goodman-Bacon, Andrew. 2021. “Difference-in-Differences with Variation in Treatment Timing.” *Journal of Econometrics* 225(2): 254–277.
- Klerman, Jacob Alex, Kelly Daley, and Alyssa Pozniak. 2012. “Family and Medical Leave in 2012: Technical Report.” Abt Associates report prepared for U.S. Department of Labor.
- Lalive, Rafael, Analía Schlosser, Andreas Steinhauer, and Josef Zweimüller. 2014. “Parental Leave and Mothers’ Careers: The Relative Importance of Job Protection and Cash Benefits.” *Review of Economic Studies* 81(1): 219–265.
- Rambachan, Ashesh, and Jonathan Roth. 2023. “A More Credible Approach to Parallel Trends.” *Review of Economic Studies* 90(5): 2555–2591.
- Rossin-Slater, Maya, Christopher J. Ruhm, and Jane Waldfogel. 2013. “The Effects of California’s Paid Family Leave Program on Mothers’ Leave-Taking and Subsequent Labor Market Outcomes.” *Journal of Policy Analysis and Management* 32(2): 224–245.
- Roth, Jonathan. 2023. “What’s Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature.” *Journal of Econometrics* 235(2): 2218–2244.

- Ruhm, Christopher J. 1998. "The Economic Consequences of Parental Leave Mandates: Lessons from Europe." *Quarterly Journal of Economics* 113(1): 285–317.
- Schönberg, Uta, and Johannes Ludsteck. 2014. "Expansions in Maternity Leave Coverage and Mothers' Labor Market Outcomes after Childbirth." *Journal of Labor Economics* 32(3): 469–505.
- Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225(2): 175–199.

A. PFL Adoption Timeline

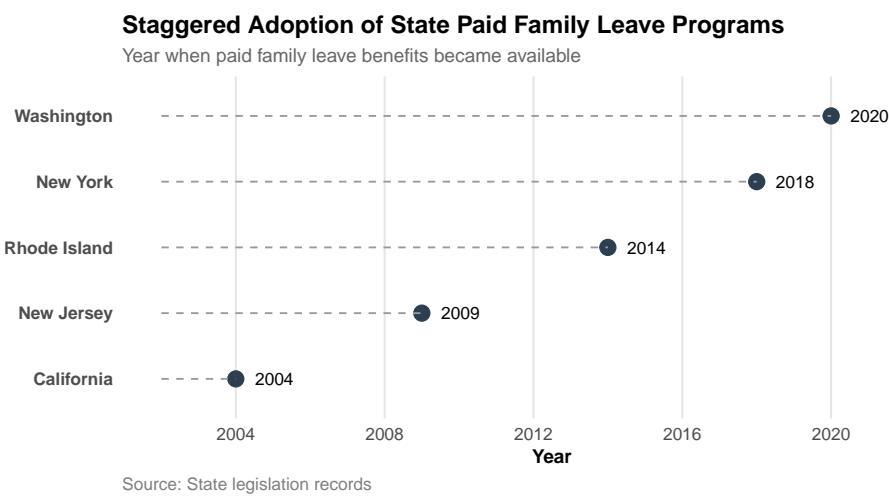


Figure 3: Staggered Adoption of State Paid Family Leave Programs

B. Heterogeneity by Education

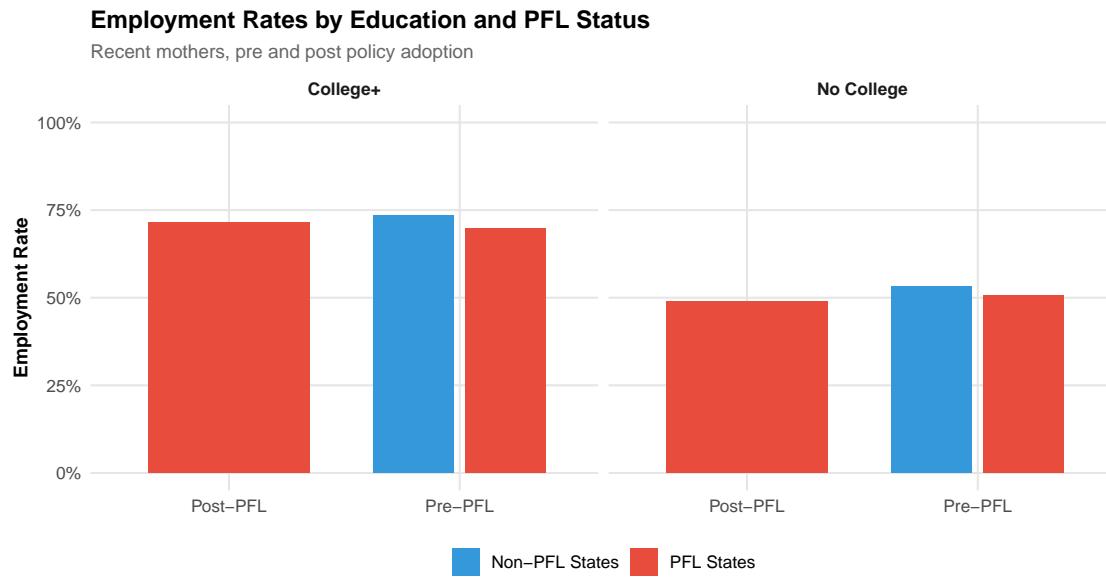


Figure 4: Employment Rates by Education and PFL Status

C. State-Specific Effects

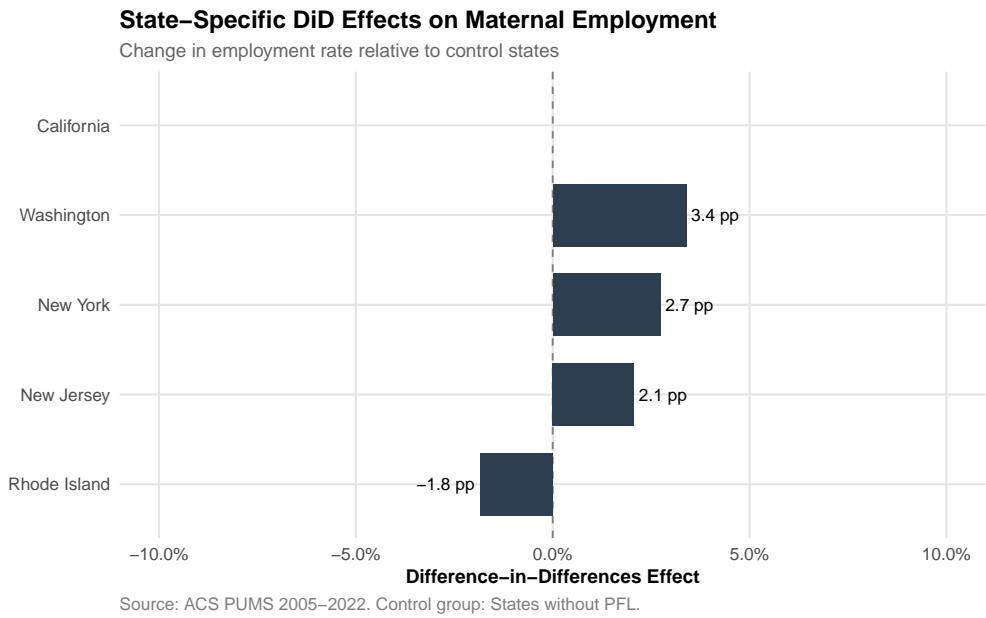


Figure 5: State-Specific DiD Effects on Maternal Employment

Notes: Simple difference-in-differences for each treated state relative to average control group change. California is excluded because it adopted PFL in 2004 (before our sample period begins in 2005), so no pre-treatment data is available. Not adjusted for heterogeneity or pre-trends.