

# Flexible Analysis of Individual Heterogeneity in Event Studies: Application to the Child Penalty\*

Dmitry Arkhangelsky<sup>†</sup>

Kazuharu Yanagimoto<sup>‡</sup>

Tom Zohar <sup>§</sup>

March 29, 2024

## Abstract

We provide a practical toolkit for analyzing effect heterogeneity in event studies. We develop an estimation algorithm and adapt existing econometric results to provide its theoretical justification. We apply these tools to Dutch administrative data to study individual heterogeneity in the child-penalty (CP) context in three ways. First, we document significant heterogeneity in the individual-level CP trajectories, emphasizing the importance of going beyond the average CP. Second, we use the individual-level estimates to examine the impact of childcare supply expansion policies. Our approach uncovers nonlinear treatment effects, challenging the conventional policy evaluation methods constrained to less flexible specifications. Third, we use the individual-level estimates as a regressor on the right-hand side to study the intergenerational elasticity of the CP between mothers and daughters. After adjusting for the measurement error bias, we find the elasticity of 24%. Our methodological framework contributes to empirical practice by offering a flexible approach tailored to specific research questions and contexts. We provide an open-source package ('unitdid') to facilitate widespread adoption.

---

\*We thank seminar participants at Carlos III, CEMFI, IFAU, and Warwick for their helpful comments. We gratefully acknowledge financial support from the Maria de Maeztu grant that funded the data access from the Dutch Central Bureau of Statistics. All views and errors are our own

<sup>†</sup>CEMFI, CEPR, [darkhangel@cemfi.es](mailto:darkhangel@cemfi.es).

<sup>‡</sup>CEMFI, [kazuharu.yanagimoto@cemfi.edu.es](mailto:kazuharu.yanagimoto@cemfi.edu.es).

<sup>§</sup>CEMFI, CESifo, [tom.zohar@cemfi.es](mailto:tom.zohar@cemfi.es), corresponding author.

# 1 Introduction

Event study analysis is ubiquitous in empirical economics. In a typical application, researchers start with assumptions that describe the behavior of a comparison group that was not affected by the event of interest and then use econometric methods, most commonly Difference-in-Differences (DiD), to quantify the effect of the event itself.<sup>1</sup> In practice, these effects tend to vary across units, and to summarize them, researchers report estimates for the average effect. However, many economic questions require us to go beyond reporting the averages and investigate the underlying heterogeneity in more detail. This is the objective we pursue in the paper.

We provide a practitioner's toolkit for analyzing unit-level effect heterogeneity in event studies where the outcomes for the comparison group follow a conditional two-way fixed-effect model.<sup>2</sup> We develop an estimation algorithm and adapt existing econometrics results to provide its theoretical justification. We apply the newly developed tools to study heterogeneity in child penalties (CP)<sup>3</sup>. Our algorithm delivers individual-level measures of the CP, which we use for three tasks. First, we document considerable observed and unobserved heterogeneity in the CP and discuss how its magnitude varies across different subpopulations. Second, we show how to use the individual-level CP estimates to investigate the economic effect of childcare provision policies on the CP. We demonstrate that our approach produces qualitatively and quantitatively different results than the policy evaluation methods currently used in the literature. Finally, we show how to use the estimated individual-level CP to study the intergenerational persistence in the CP.

Our proposal is based on a simple observation that has a long history in the econometric panel data literature (Chamberlain, 1992; Bonhomme and Sauder, 2011; Arellano and Bonhomme, 2012). Suppose we compare a given unit before and after the event to an average of appropriate control units. As long as we expect the parallel trends assumption to hold, this comparison delivers an unbiased, though noisy, estimator for unit-level effects. We recommend that researchers directly use these unit-specific measures in the regression analysis, either as a left-hand side (LHS) variable, which is relevant for studying heterogeneity with respect to observables and conducting policy evaluation exercises, or as a right-hand side (RHS) variable, which is applicable, for example, in intergenerational questions.

One immediate concern with our recommendation is that the constructed unit-specific measures are noisy, and the regression-based analysis based on such variables can potentially

---

<sup>1</sup>Often, researchers interpret the result of this exercise as causal evidence, but our analysis does not rely on this interpretation.

<sup>2</sup>See our publically available R package at: <https://github.com/kazuyanagimoto/unitdid>

<sup>3</sup>In their influential paper, Kleven et al. (2019b) coined the term "child penalty" to describe the differential career and earnings losses incurred by women compared to men after having children.

suffer from a measurement error bias. We show that this bias does not appear when we put the estimates on the LHS and use them for policy analysis as long as we restrict attention to policies that vary across prespecified subpopulations of units and periods. For example, this is the case in our empirical application, where the childcare provision policy varies geographically. Intuitively, when analyzing such policies, we need to aggregate unit-level estimates to the level of the policy variation, and this aggregation eliminates the measurement error.

The situation is different when we put the noisy estimates on the RHS, with the measurement error leading to the familiar attenuation bias. We show how to adjust for this bias by using the information from the comparison group. This adjustment relies on a generalization of the parallel trends assumption. In particular, we require the means and variances of the appropriately transformed outcomes to be stable across event times. Similar to the conventional parallel trends, this assumption has testable implications, and we validate it in the context of our empirical application.

We apply our method to study CP in the Netherlands. We utilize a comprehensive dataset from Statistics Netherlands (CBS), encompassing various sources, including municipality registers and tax records, all linked through anonymized identifiers. Our primary data comes from employer-employee monthly records from 1999 to 2016, detailing employment statuses and wages. We merge demographic information about birth year, gender, timing of parenthood, and education records on the highest diploma obtained.

We start our empirical analysis by constructing the individual-specific CP estimates. We conduct a naive exploratory analysis to understand the variation in estimates, ignoring measurement errors. Clustering individuals into three groups based on their estimated CP trajectories reveals significant differences in labor market trajectories after childbirth for both fathers and mothers. These findings suggest averaging CP across the entire population may obscure substantial individual-level differences. Observable individual characteristics and the timing of childbirth can explain part of the variation in the estimated CP. To investigate this dimension, we project the individual-level estimates on birth cohort and age at first childbirth fixed effects. We find that while the CP diminishes over successive birth cohorts and with older ages at first childbirth, the variability within these groups underscores a critical selection effect tied to the timing of parenthood.

We continue our analysis by investigating the role of childcare provision in labor supply by examining childcare service data and availability across municipalities. We focus on labor market outcomes surrounding the first childbirth from 2000 to 2022, aiming to assess the impact of a 2005 childcare provision expansion policy on the CP. We find a nonlinear treatment effect of childcare provision expansion on almost all margins, groups, and years post-birth. For example, the most striking nonlinear effects are evident for the fathers' labor-force participation. These U-shaped effects suggest that an expansion of childcare has a greater impact

on the father's labor supply when the baseline availability of childcare is low. However, in higher levels of baseline childcare, a childcare expansion can also improve father's labor-force participation.

These childcare provision expansion policies have been recently studied, finding a wide range of empirical evidence (Rabaté and Rellstab, 2021; Andresen and Nix, 2022a; Kleven et al., 2022; Lim and Duletzki, 2023; Castellanos, 2024). Compared to the standard analysis in the literature, we approach this question in a much more flexible way, again demonstrating the strength of our method. Typically, both the CP and their reaction to policies are estimated within a single regression which is augmented with pre-post policy indicators interacted with binary measures of exposure. Such specifications raise several concerns. First, in practice, the effect of the policy is often mediated through a continuous measure (e.g., childcare provision). In this case, binary exposures ignore important information that we can directly use. This is particularly important if we suspect the effect of the policy is nonlinear. Second, standard specifications abstract away from dynamic effects, which can be important in applications. Finally, these specifications suffer from aggregation issues highlighted recently in the literature on multiple treatments (Goldsmith-Pinkham et al., 2022).

In theory, these issues can be addressed by specifying a regression model that incorporates dynamics and nonlinearity and using an estimation method that accounts for potential aggregation issues. In practice, there are several advantages in separating the two exercises – construction of unit-level estimates and policy evaluation – rather than implementing them in a single step. For example, applied researchers often consider different specifications in their policy evaluation exercises. By doing the analysis in two steps, we can guarantee that the type of policy evaluation exercise we implement does not interfere with the way we measure the unit-level effects. Moreover, some policy evaluation techniques rely on statistical methods that involve regularization and nonlinearities. Such methods are part of the standard toolkit of an empirical economist when the outcomes are readily available, and our approach allows researchers to use the same tools in environments with constructed outcomes.

Next, we turn to the final part of our empirical analysis, which focuses on using the estimated CP as a regressor. As discussed above, this is not straightforward because we need to account for the measurement error bias. We approach this task in two steps. First, we rely on the richness of the Dutch data to estimate the conditional variance in the CP and investigate the validity of the generalized parallel trends assumption on which our analysis relies. The data supports this assumption, at least for relatively older individuals with a well-established labor market history. The results of this exercise are also valuable on their own because they allow us to uncover the magnitudes of the unobserved heterogeneity (as quantified by its variance) and its variation across observed characteristics.

Unfortunately, we cannot use the Dutch data to investigate intergenerational-mobility (IGM)

questions because we observe outcomes only for two decades, which is insufficient to estimate the CP for parents and kids. Therefore, we turn to PSID and find a naive CP-IGM coefficient of 0.148 and the bias-corrected coefficient of 0.236, implying a 10% increase in mothers' CP is associated with a 2.36% higher CP for the daughters. This result suggests an economically important role for the transmission of social norms between generations in the U.S. However, it is important to take these specific estimates with a grain of salt, given the small sample size of the PSID.

Our paper makes both a methodological and an empirical contribution. On the methodological side, we combine the available results from the econometric panel data literature (Arellano and Bonhomme, 2012) and recent literature on event studies (Borusyak et al., 2023) to develop a simple yet highly flexible algorithm that can be applied to all event studies. In Arellano and Bonhomme (2012), the authors show how to use heteroskedasticity restrictions to identify the variance of the random coefficients in a linear panel data model. Our identification results extend their analysis to environments where the dimension of unobserved heterogeneity is so large that it cannot be recovered in an unbiased way for any given unit in the data. Yet, we can recover parts of it for specific units. We also show how to use the (conditional) variance to solve the measurement error problem, thus allowing researchers to use estimate unit-level effects as regressors. We extend the analysis in (Borusyak et al., 2023) by constructing unit-specific variance measures that play a dual role: they allow us to investigate the magnitudes of unobserved heterogeneity and adjust for the measurement error bias. Our two-step approach, where researchers first construct unit-level outcomes and then analyze them, has broad implications for empirical practice. For example, by separating the measurement of the effects from policy evaluation, our methodology facilitates a more flexible analysis that can be tailored to specific research questions and contexts.

The idea of creating unit-level measures based on appropriately transformed data has a long tradition in econometrics and causal inference. The early work of Robinson (1988) showed how to use appropriately transformed outcomes and regressors to construct efficient estimators in partially linear models. More recently, these ideas were extended in the causal inference literature with applications to heterogenous treatment effects and thier summaries (Chernozhukov et al., 2018, 2023; Kennedy, 2023; Nie and Wager, 2021; Semenova and Chernozhukov, 2021), and policy learning (Athey and Wager, 2021). See Foster and Syrgkanis (2023) for a comprehensive treatment and many additional references. Our proposal not only brings similar ideas to event-study settings but also generalizes them by allowing researchers to use unit-specific measures as RHS variables. The key limitation of our analysis compared to this literature is the focus on discrete covariates.

On the empirical side, this paper makes several contributions to the CP literature. It sheds light on the heterogeneity of the CP across individuals, complementing and expanding upon

a body of research that has predominantly focused on aggregate penalties. Previous studies have consistently shown that women endure a substantial CP post-birth, from which recovery is rare if not non-existent (Angelov et al., 2016; Kleven et al., 2019b,a; Gallen, 2019). In response to this gap, our empirical investigation introduces an innovative individual measure of the CP, which allows us to document and analyze the high heterogeneity in the CP across different dimensions, including the mean and variance of the penalty. This approach reveals a complex landscape of the CP that challenges the standard aggregate analyses.

Using these individual CP estimates, we provide the *first* estimates of the IGM in CP as a measure for the transmission of social norms between generations. Kleven et al. (2019b) provide some estimates of an intergenerational transmission of CP, by plotting the mothers' CP in earnings against quintiles of the relative labor supply distribution of the grandparents. Our methodology allows us to go beyond and use estimates of the CP in earnings for both generations.

This narrative is further nuanced by research exploring the effects of childcare expansion policies on the CP, with mixed evidence across different national contexts, including Norway (Andresen and Nix, 2022b), Austria (Kleven et al., 2022), Germany (Lim and Duletzki, 2023), Canada (Karademir et al., 2024), and Spain (Castellanos, 2024). These studies collectively highlight the varied efficacy of policy interventions in mitigating the CP, underscoring the need for a more granular analysis to capture the underlying dynamics. Our findings particularly underscore the non-monotonic impact of childcare policies on women's and men's labor market choices and composition. By employing this individual-level measure, we can dissect the nuances of policy impacts, revealing that while some childcare expansion initiatives have been successful in reducing the CP to a degree, their effectiveness is highly contingent on a range of factors, such as the baseline levels of childcare before the expansion.

Notably, contemporary work by Castellanos (2024) uses the estimated individual CP for policy analysis. Our main distinctions relative to Castellanos (2024) are three-fold: (1) formalizing a general algorithm for this estimation procedure that extends beyond the the CP to other event studies; (2) providing a method to use these individual estimates as regressors; (3) quantifying the magnitudes of the unobserved heterogeneity. Finally, we provide an open-source package for practitioners ('unitdid').

This detailed examination of the CP heterogeneity enriches the existing literature by providing a more nuanced understanding of the effects of childbirth and childcare policies on women's labor market outcomes and offers critical insights for policymakers. The revelation of a non-monotonic impact of childcare policies underscores the importance of considering the diverse needs and circumstances of different population segments when designing interventions to mitigate the CP. Through this empirical contribution, we highlight the significance of adopting a more individualized approach to understanding and addressing the challenges

women face in the labor market post-childbirth, paving the way for more effective and equitable policy solutions.

The rest of this article is organized as follows. In Section 2, we formally introduce our setup and main assumptions. Section 3 describes the data and sample selection and the application of our method to quantify the extent of unobserved heterogeneity in the CP. Section 4 explains how the unit-level estimates can be used as an outcome of interest on the LHS in policy evaluation exercises and applies the developed methodology to evaluate childcare provision policies. Section 5 explains how to account for the measurement error in estimated unit-level effects so that they can be used as a RHS variable. Section 6 concludes.

## 2 Econometric framework

In this section, we formally introduce our setup and key assumptions, which we maintain throughout the paper. In the later sections, we discuss additional restrictions needed to identify specific objects. The identification arguments we use are well-understood from the point of view of the econometrics panel data literature (Chamberlain, 1992; Arellano and Bonhomme, 2012), and we present them largely to introduce the notation and set the stage for the analysis in the rest of the paper.

### 2.1 Setup

The sample consists of  $n$  units, with  $i$  denoting a generic one, which we observe over calendar time  $t \in \mathcal{T}_i := \{T_{0,i}, \dots, T_{1,i}\}$ . For each unit  $i$ , we observe outcomes  $Y_{i,t}$  in period  $t$  and fixed characteristics  $X_i$ .<sup>4</sup> In what follows we treat  $\mathcal{T}_i$  as a part of  $X_i$ . Each unit is characterized by an event time  $E_i^* \in \mathcal{E}$ , and we observe a right-censored version of this variable:

$$E_i = \begin{cases} E_i^*, & \text{if } E_i^* \leq T_{1,i} \\ +\infty, & \text{if } E_i^* > T_{1,i}. \end{cases}$$

We view the units as an i.i.d. sample from a population of interest and collect the observed data into a dataset  $\mathcal{D}$ .

Our first assumption describes the model for the outcomes.

#### Assumption 2.1. (*Two-Way Model*)

---

<sup>4</sup>Time-varying covariates  $X_{i,t}$  can be incorporated by combining them into a tuple  $X_i := \{X_{i,T_{0,i}}, \dots, X_{i,T_{1,i}}\}$ .

For some known  $K \geq 0$  and each  $t \in \mathcal{T}_i$  we have

$$Y_{i,t} = \alpha_i + \lambda_t(X_i) + \sum_{k \geq -K} \tau_{i,e,k} \mathbf{1}\{E_i^* = e, t - e = k\} + \epsilon_{i,t},$$

$$\mathbb{E}[\epsilon_{i,t}|X_i, \alpha_i, \boldsymbol{\tau}_i] = 0,$$

where  $\boldsymbol{\tau}_i := \{\tau_{i,e,k}\}_{e \in \mathcal{E}, k \geq -K}$ .

We can interpret Assumption 2.1 as a model for realized values of the underlying potential outcomes in which the benchmark outcomes follow a (conditional) two-way model, and there are limited anticipation effects for at most  $K$  periods; for a recent survey of such models, see Arkhangelsky and Imbens (2023). Alternatively, we can interpret the same structure as a statistical model that connects the latent variables,  $\boldsymbol{\tau}_i$ , to observed outcomes and event times.

To understand this model, consider a unit  $i$  with the adoption date  $E_i^*$ . Suppose  $E_i^* \leq T_{1,i}$  and thus  $E_i^* = E_i$ , i.e., the observed event time is not censored. Then, for any period  $t \in \mathcal{T}_i$  prior to  $E_i - K$ , the evolution of the outcomes for this unit is determined by the two-way model:

$$Y_{i,t} = \alpha_i + \lambda_t(X_i) + \epsilon_{i,t}. \quad (1)$$

At the same time, for any period  $t \in \mathcal{T}_i$  such that  $t \geq E_i - K$ , the evolution of the outcomes is shifted by  $\tau_{i,E_i,t-E_i}$ :

$$Y_{i,t} = \alpha_i + \lambda_t(X_i) + \tau_{i,E_i,t-E_i} + \epsilon_{i,t}. \quad (2)$$

The situation is more challenging for units for which the event time is censored; that is, we only know that  $E_i^* > T_{1,i}$ . Then for any  $t \in \mathcal{T}_i$  such that  $t \leq T_{1,i} - K$  we have the following:

$$Y_{i,t} = \alpha_i + \lambda_t(X_i) + \epsilon_{i,t},$$

but the model for the outcomes in period  $t \in \{T_{1,i} - K + 1, \dots, T_{1,i}\}$  cannot be written in terms of observed quantities and unit-specific parameters introduced in Assumption 2.1. We discard outcomes in these periods because they do not contain information we can use for identification.

Define  $\boldsymbol{\epsilon}_i^\top := (\epsilon_{i,T_{0,i}}, \dots, \epsilon_{i,T_{1,i}})$ ; the following assumption restricts the data generating process behind  $E_i^*$ .

**Assumption 2.2. (STRICT EXOGENEITY)**

Adoption time  $E_i^*$  is strictly exogenous with respect to  $\boldsymbol{\epsilon}_i$ :

$$\boldsymbol{\epsilon}_i \perp\!\!\!\perp E_i^* | X_i, \alpha_i, \boldsymbol{\tau}_i.$$

To unpack this assumption, it is instructive to compare it to an experimental benchmark. Suppose the event time  $E_i^*$  was assigned in a completely randomized experiment, with the resulting outcomes being generated according to Assumption 2.1. In this case, Assumption 2.2 trivially holds by design. The same logic applies to more complicated setups where we can view  $E_i^*$  as quasi-random, with a distribution that varies by  $X_i$ . Assumption 2.2 is more general and allows the assignment to be based on the unobserved unit-specific parameters, including the whole trajectory  $\tau_i$ . In particular, if we interpret 2.1 as a causal model, then this allows for the selection based on future gains. As a result, it is a fairly general restriction that we can expect to hold in a large class of applications.

**Remark 2.1.** The model specified by Assumptions 2.1 - 2.2 has a user-specified parameter  $K$ , which quantifies the extent of anticipation analogously to the limited-anticipation assumption in Callaway and Sant'Anna (2021). In practice, we choose this parameter by analyzing the behavior of the outcomes prior to event time  $E_i$ .

**Remark 2.2.** The two-way model in Assumption 2.1 allows the benchmark outcomes to vary freely with  $X_i$  but restricts the systematic variation within the subpopulations with the same value of  $X_i$ . We focus on this model given its prevalence in applied work, but our approach can be naturally extended to environments with interactive fixed effects (e.g., Chamberlain, 1992; Arellano and Bonhomme, 2012). In some applications (Kleven et al., 2019a), researchers consider models without unit-fixed effects  $\alpha_i$ , which is a special case of Assumption 2.1.

## 2.2 Sample selection

In our analysis, we restrict attention to a particular subsample of units. This subsample depends on another user-specified parameter (in addition to  $K$ ), which we denote by  $H$ . This parameter describes the horizon we are analyzing; in our empirical applications, we always set  $H = 5$ . Formally, most variables that we define from now on should be indexed by  $K$  and  $H$ , but we omit this indexing for brevity.

First, we define a set of units we will use for the analysis. To be able to eliminate unit-level fixed effects  $\alpha_i$ , we need to restrict attention to event times that are sufficiently distant from the starting observation period. We define the appropriate selection indicator,

$$S_i := \mathbf{1}\{T_{0,i} + K < E_i^*\},$$

and denote the set of such units by  $\mathcal{S}$ , i.e,  $\mathcal{S} := \{i \in [n] : S_i = 1\}$ . Next, we define the selected time periods:

$$\mathcal{T}_i^s := \begin{cases} \{T_{0,i}, \dots, T_{1,i}\}, & \text{if } E_i \leq T_{1,i} \\ \{T_{0,i}, \dots, T_{1,i} - K\}, & \text{if } E_i > T_{1,i} \end{cases}$$

In our analysis, we focus on units in  $\mathcal{S}$  and for any such unit, we restrict attention to periods  $t \in \mathcal{T}_i^s$ .

Units in  $\mathcal{S}$  serve different purposes, and we split this set accordingly. First, we define a set of units that we call “treated”:

$$S_i^{tr} := \mathbf{1}\{S_i = 1; E_i \leq T_{1,i} - H\},$$

and collect them into a set  $\mathcal{S}^{tr}$ .<sup>5</sup> For every unit  $i \in \mathcal{S}^{tr}$ , we observe outcomes for at least  $H$  periods after the event time. We further split  $\mathcal{S}$  and  $\mathcal{S}^{tr}$  based on event times. In particular, for any  $e \in \mathcal{E}$  we define:

$$\begin{aligned} S_i^{tr}(e) &:= \mathbf{1}\{E_i = e; S_i = 1; E_i \leq T_{1,i} - H\}, \\ S_i^{cont}(e) &:= \mathbf{1}\{E_i > e; S_i = 1\}; \end{aligned}$$

The first indicator describes units in  $\mathcal{S}$  with event time  $e$  for which we observe data at least  $K + 1$  periods before  $e$  (because  $S_i = 1$ ) and at least  $H$  periods after. The second indicator describes units in  $\mathcal{S}$  with event time larger than  $e$ . We use  $\mathcal{S}^{cont}(e)$  and  $\mathcal{S}^{tr}(e)$  to denote the corresponding subpopulations for a given event time  $e$ .

## 2.3 Unit-specific measures

Equation (2) shows that for any  $t \geq E_i - K$ , the observed outcomes are informative about  $\tau_{i,E_i,t-E_i}$ , as long as we account for the bias introduced from a unit-specific unobserved heterogeneity that is fixed across time ( $\alpha_i$ ), and for the time-specific effects that might vary by observables of interest ( $\lambda_t(X_i)$ ). Intuitively, one can recover these objects by estimating equation (1) by ordinary least squares (OLS) using units in  $\mathcal{S}^{cont}$  and pre-event periods  $t < E_i - K$ . This intuition masks a few subtleties in the estimation process. Therefore, we opt to describe the estimation process formally.

For any unit  $i \in \mathcal{S}^{cont}(e + k + K)$  such that  $e + k \in \mathcal{T}_i^s$  we define:

$$\begin{aligned} \Delta Y_{i,e,k} &:= Y_{i,e+k} - \frac{\sum_{T_{0,i} \leq l < e-K} Y_{i,l}}{e - T_{0,i} - K} = \\ &\quad \left( \lambda_{e+k}(X_i) - \frac{\sum_{T_{0,i} \leq l < e-K} \lambda_l(X_i)}{e - T_{0,i} - K} \right) + \left( \epsilon_{i,e+k} - \frac{\sum_{T_{0,i} \leq l < e-K} \epsilon_{i,l}}{e - T_{0,i} - K} \right). \end{aligned}$$

Assumptions 2.1 - 2.2 guarantee that  $\Delta Y_{i,e,k}$  provides an unbiased signal for a particular com-

---

<sup>5</sup>Note that “treated” units in  $\mathcal{S}^{tr}$  can still serve as “control” observations for units with earlier adoption dates.

bination of time-specific functions:

$$\Delta\lambda_{e,k}(x) := \left( \lambda_{e+k}(x) - \frac{\sum_{T_0(x) \leq l < e-K} \lambda_l(x)}{e - T_0(x) - K} \right) = \mathbb{E}[\Delta Y_{i,e,k}|X_i = x, S_i^{cont}(e+k+K) = 1, e+k \in \mathcal{T}_i^s]. \quad (3)$$

We use  $\Delta\lambda_{e,k}(x)$  evaluated at  $(E_i, X_i)$  to define for  $i \in \mathcal{S}^{tr}$  and  $k \in \{-K, \dots, H\}$ :

$$\tau_{i,k} := \Delta Y_{i,E_i,k} - \Delta\lambda_{E_i,k}(X_i) = \tau_{i,E_i,k} + \left( \epsilon_{i,E_i+k} - \frac{\sum_{T_{0,i} \leq l < E_i-K} \epsilon_{i,l}}{E_i - T_{0,i} - K} \right).$$

In practice, we cannot compute expectations in (3) and instead rely on noisy estimates  $\hat{\Delta}\lambda_{e,k}(x)$ . Algorithm 1 summarizes this approach, defining  $\hat{\Delta}\lambda_{e,k}$  as a solution to an OLS optimization problem over some functional class  $\mathcal{F}$  without specifying the exact nature of this class. In our empirical application,  $X_i$  is discrete, and thus  $\mathcal{F}$  includes all possible functions of  $x$ , which corresponds to computing  $\hat{\Delta}\lambda_{e,k}(x)$  using empirical analogs of (3). In addition to  $\hat{\tau}_{i,k}$ , the algorithm delivers  $\hat{Y}_{i,k}^c$ , which we later use as estimates of the benchmark outcomes,

$$Y_{i,k}^c := \alpha_i + \lambda_{E_i+k}(X_i) + \frac{\sum_{T_{0,i} \leq l < E_i-K} \epsilon_{i,l}}{E_i - T_{0,i} - K}.$$

Algorithm 1 uses as inputs data  $\mathcal{D}$ , two user-specified parameters  $K$  and  $H$ , the functional class  $\mathcal{F}$ , and weights  $\{\omega_i\}_{i \in \mathcal{S}}$ . These weights are needed to conduct inference, and they affect the output of the algorithm only through the construction of the function  $\hat{\Delta}\lambda_{e,k}(x)$ . We construct them by generating a sample of i.i.d. random variables  $\{\omega_i\}_{i \in \mathcal{S}}$ , with  $\omega_i \sim \text{Exp}(1)$ . To conduct inference, we rerun Algorithm 1 many times, each time generating a new set of weight  $\{\omega_i\}_{i \in \mathcal{S}}$ . The collection of resulting estimates can then be used for inference using standard Bayesian bootstrap (Chamberlain and Imbens, 2003).

**Remark 2.3.** In Borusyak et al. (2023), the authors use OLS-based imputations to construct unit-specific measures of treatment effects, which are later aggregated with user-specified weights in order to produce the average effect of interest. We choose to construct unit-specific measures using Algorithm 1, because it delivers an explicit expression for the measurement error, which we use in later sections.

**Remark 2.4.** In our discussion, we ignored the support conditions, i.e., we assumed that expectations in (3) are well-defined for any  $(x, e)$  in the support of  $(X_i, E_i)|S_i^{tr} = 1$ , allowing us to use  $\Delta\lambda_{E_i,k}(X_i)$  to construct  $\tau_{i,k}$ . In practice, the set  $\mathcal{S}^{cont}(e+k+K)$  can be empty for some relevant values of  $e$ . In this case, we drop such event times from the analysis, thus introducing an additional sample selection restriction.

**Algorithm 1:** Outcome construction

```

Data:  $\mathcal{D}, K, H, \mathcal{F}, \{\omega_i\}_{i \in \mathcal{S}}$ 
Result:  $\{\hat{Y}_{i,k}^c, \hat{\tau}_{i,k}\}_{i \in \mathcal{S}^{tr}; k \in \{-K, \dots, H\}}$ 
1 for  $e \in \mathcal{E}$  do
2   for  $k \in \{-K, \dots, H\}$  do
3     for  $i : E_i \geq e > T_{0,i} + K$  and  $e + k \in \mathcal{T}_i^s$  do
4       Define  $\Delta Y_{i,e,k} := Y_{i,e+k} - \frac{\sum_{T_{0,i} \leq l < e-K} Y_{i,l}}{e - T_{0,i} - K}$ 
5     end
6     Define
7       
$$\Delta \hat{\lambda}_{e,k} := \arg \min_{f \in \mathcal{F}} \sum_{i \in \mathcal{S}(e+k+K), e+k \in \mathcal{T}_i^s} (\Delta Y_{i,e,k} - f(X_i))^2 \omega_i$$

8     for  $i \in \mathcal{S}^{tr}(e)$  do
9       Define  $\hat{\tau}_{i,k} := \Delta Y_{i,e,k} - \Delta \hat{\lambda}_{e,k}(X_i)$ ;
10      Define  $\hat{Y}_{i,k}^c := \Delta \hat{\lambda}_{e,k}(X_i) + \frac{\sum_{T_{0,i} \leq l < e-K} Y_{i,l}}{e - T_{0,i} - K}$ ;
11    end
12  end
13 end

```

## 2.4 Measurement error

We will use estimated  $\hat{\tau}_{i,k}$  both as left-hand-side (LHS) and right-hand-side (RHS) variables, and to understand the problems this leads to, we write it in the following form:

$$\hat{\tau}_{i,k} = \tau_{i,k} + \xi_{i,k} = \tau_{i,E_i+k} + \left( \epsilon_{i,E_i+k} - \frac{\sum_{T_{0,i} \leq l < E_i - K} \epsilon_{i,l}}{E_i - T_{0,i} - K} \right) + \xi_{i,k},$$

where  $\xi_{i,k}$  is the estimaton error in  $\hat{\tau}_{i,k}$  that arises from the error  $\hat{\lambda}_t(X_i) - \lambda_t(X_i)$ . Under weak conditions on the underlying estimator this error vanishes as  $n$  goes to infinity. The second error term,

$$\epsilon_{i,k} := \left( \epsilon_{i,E_i+k} - \frac{\sum_{T_{0,i} \leq l < E_i - K} \epsilon_{i,l}}{E_i - T_{0,i} - K} \right), \quad (4)$$

is unit-specific and does not vanish with  $n$ . We refer to it as the measurement error (ME) in  $\hat{\tau}_{i,k}$  for  $\tau_{i,E_i+k}$ . Assumptions 2.1 - 2.2 guarantee that the ME is orthogonal to  $X_i$ :

$$\mathbb{E} [\epsilon_{i,k} | X_i, \alpha_i, \boldsymbol{\tau}_i, S_i^{tr} = 1] = 0. \quad (5)$$

This condition implies that up to vanishing errors  $\xi_{i,k}$ , we can safely use  $\hat{\tau}_{i,k}$  as a LHS variable for the regression analysis instead of  $\tau_{i,E_i+k}$ .

### 3 Heterogeneity in child-penalty

In this section, we implement Algorithm 1 for a particular empirical problem – estimation of child penalties. In their influential paper, Kleven et al. (2019b) coined the term “child penalty” (CP) to describe the differential career and earnings losses incurred by women compared to men after having children. We begin by describing the data we use for the CP estimation. We then describe the estimation of individual-level CP and explore the observed and unobserved heterogeneity in CP. These data and individual-level estimates are then used in Section 4, where we analyze the effects of childcare expansion policies on the CP.

#### 3.1 Data

##### Data Sources

We use administrative data from Statistics Netherlands (CBS) on the universe of Dutch residents. Different data sources, such as the municipality registers or tax records, are matched through unique individual or household anonymized identifiers. In the following section, we present the main variables used and the sample construction.

**Tax and Employment Records** Our primary source of data is an extensive yearly-level employer-employee data set derived from tax records (baansommentab), covering the period from 1999 to 2016. We analyze two labor market outcomes: unconditional earnings and employment. Employment is specified as having a job based on an employment contract between a firm and a person, excluding self-employment. Second, the earnings data consists of yearly gross earnings after social security contributions, but before taxes and health insurance contributions from official tax data.

**Demographic and Education Information** To enrich our understanding of the workforce, we incorporate demographic data into our analysis (gbapersoontab). This includes information such as birth year, date of death, sex, and yearly information on the municipality of residence, household composition, marital status, and migration spells (gbaadresobjectbus and vslgwbtab).

A unique aspect of our demographic data is the inclusion of a parent-child key (kindoudertab). We use information on birth dates and linkage between parents and their children to determine the first child for all legal parents, which may include both adoptive or biological parents.

Lastly, we also observe the educational attainment at each point in time (hoogsteopltab) and use the highest level of education attained by 2022.

**Childcare Provision Data** An integral part of our study involves examining the role of childcare in labor market participation. To this end, we use records about childcare service providers using the firm’s job classification (betab), and data on job location that we use to compute our index of childcare supply per municipality (gemstplaatsbus/gemtplbus/ngemstplbus). The job location data set contains the municipality of each job, measured in December. When there is more than one establishment for a given firm, we impute the closest establishment to the worker’s residential address as the job location.

### Sample Definition

A key aspect of our study is the examination of labor market outcomes around the time of first childbirth. We specifically look at individuals who had their first child between 2003 and 2022. Furthermore, we restrict our sample to individuals that are 22 to 45 years old at the time of first birth, a critical phase for career development and family planning.

Note that in Section 4 we study the effect of a 2005 childcare provision expansion policy on the child-penalty. Therefore, we adjust our sample criteria to include parents with their first child-birth from 2000 so that we can see the full horizon of their labor market five year post birth before the 2005 policy took place.

## 3.2 Exploring heterogeneity

To estimate individual CP we follow the approach outlined in Section 2. We restrict attention to three different covariates: gender  $G_i \in \{\text{male; female}\}$ , education level  $educ_i$ , and birth cohort  $B_i$ , so that  $X_i = (G_i, educ_i, B_i)$ . We apply Algorithm 1 with  $H = 5$  and  $K = 3$ , and since  $X_i$  is discrete, we estimate  $\Delta\hat{\lambda}_{e,k}(x)$  using the empirical analog of (3).

We use the resulting data set to construct estimates of the individual CP. Since CP are more naturally interpreted in relation to underlying baseline outcomes (earnings or participation), we use measures  $\hat{Y}_{i,k}^c$  also delivered by Algorithm 1 to normalize them. In particular, consider an individual  $i \in S^{tr}$  with  $X_i = x$  and  $E_i = e$ , and let  $n_{x,e}$  be the total number of such individuals. For every  $k \in \{-K, \dots, H\}$  we define

$$\tilde{\tau}_{i,k} := \frac{\hat{\tau}_{i,k}}{\frac{\sum_{j:E_j=E_i, X_j=X_i} \hat{Y}_{j,k}^c}{n_{x,e}}}, \quad (6)$$

where  $\hat{\tau}_{i,k}$  and  $\hat{Y}_{i,k}^c$  are constructed using Algorithm 1. As discussed in Section 2, the numerator of (6) is an estimator of the individual level CP  $\tau_{i,E_i,k}$ . The denominator captures the average benchmark outcome for individuals in the appropriate sub-population. For future reference,

we collect the normalized estimated coefficients into a  $(K + H + 1)$ -dimensional vector:

$$\tilde{\tau}_i := (\tilde{\tau}_{i,-K}, \dots, \tilde{\tau}_{i,H}).$$

We proceed with an exploratory analysis of these objects to visualize the variation in  $\tilde{\tau}_{i,k}$ . This analysis intentionally ignores the errors in  $\tilde{\tau}_{i,k}$  and  $\hat{Y}_{i,k}^c$ , which we discussed in Section 2. We start by plotting the marginal distributions of  $\tilde{\tau}_{i,k}$  (for earnings) across different horizons and genders pooled across all birth cohorts. The results are reported in Figure Ia and demonstrate a large variation in estimated individual CP, which increases over  $k$ . As we will see later, a large part of this variation is explained by the birth cohort  $B_i$  and the age of giving birth  $A_i := E_i - B_i$ .

Figure Ia focuses on marginal distributions for each  $k$ , and thus cannot speak to any persistence of individual CP. To address this, we apply a K-means algorithm (Lloyd, 1982) to the vector of estimated child penalties  $\tilde{\tau}_i$  and classify all individuals of the same gender into three groups. The results of this analysis are reported in Figure Ib and show important variation in the CP trajectories. In particular, we see that for around two-thirds of the male population, CP are non-existent (orange line). At the same time, a quarter of the sample have large negative trajectories (green), and another 10% of the population with a positive trajectory compared to their counterfactual earnings growth in the absence of parenthood (blue). The situation for females is qualitatively similar, though the trajectories themselves look very different, with 28% of women that seem to leave the labor force (green), 55% of the female shift to some part-time (orange), and 17% of women maintaining their pre-birth trajectory. These results suggest that the conventional analysis that reports the average CP across the entire population may be missing important individual-level variation.

As our next exercise, we summarize the heterogeneity in CP that is related to observables. We do this by estimating the following equations by OLS separately for men and women:

$$\begin{aligned}\tilde{\tau}_{i,k} &= \sum_{b=1978}^{1984} \beta_{b,k} \mathbf{1}\{B_i = b\} + \nu_{i,k}^{(b)}, \\ \tilde{\tau}_{i,k} &= \sum_{a=25}^{36} \gamma_{a,k} \mathbf{1}\{A_i = a\} + \nu_{i,k}^{(a)}.\end{aligned}$$

Figure A.1 reports the estimation results for all horizons. On a first appearance, the second figure suggests that the variation in birth cohorts is not important. This, however, is a composition effect: most individuals that we observe from 1978 are relatively old by the time we start observing their outcomes, and the opposite is true for those born in 1984. In contrast, the results in Figure A.1a show a wide variation in average child penalties by age of first childbirth,

suggesting selection into the age of giving birth plays an important role, and emphasizing the role of individual estimation.

To address the composition concern, we do a two-way decomposition of the average (over time) individual level CP by estimating the following equation by OLS separately for men and women:

$$\bar{\tau}_i = \sum_{b=1978}^{1983} \beta_b \{B_i = b\} + \sum_{a=25}^{36} \gamma_a \{A_i = a\} + \nu_i,$$

where  $\bar{\tau}_i := \frac{1}{K+H+1} \sum_{k \geq -K}^H \tilde{\tau}_{i,k}$ .

The results are reported in Figure II. Note that we omit the birth cohort of 1984, so all coefficients should be interpreted with respect to that benchmark. We see the share of variation in CP across birth cohorts and age at childbirth is similar in nature and magnitude – the CP is shrinking over birth cohorts and age of childbirth. However, it is important to note that the variation in the age of childbirth (seven ages) seems to be twice as important as the variation in birth cohort (15 birth cohorts).

Finally, the results in Figure A.1a serve as a diagnostic test for the limited-anticipation assumption. We can see a confirmation of the limited-anticipation assumption for individuals who became parents in their thirties (i.e., the estimates three periods before birth are close to zero), while this is not the case for individuals who became parents in their late twenties. This diagnostic test suggests one should focus the analysis on individuals who became parents in their thirties, a sample selection we add to the rest of our analysis.

## 4 Using individual estimates as an outcome (LHS)

In this section, we continue with our CP analysis and explain how the individual-level estimates described in Section 3 can be used as an outcome of interest in policy evaluation exercises. Specifically, we illustrate it in the context of evaluating the effects of childcare supply expansion policies on the child penalty, a policy that had been recently studied, finding a wide range of empirical evidence (Rabaté and Rellstab, 2021; Andresen and Nix, 2022a; Kleven et al., 2022).<sup>6</sup> We start by formalizing the policy analysis within the econometric framework described in Section 2, and introducing additional assumptions. We then provide the necessary institutional background on these policies, show how to use our measures to analyze them,

---

<sup>6</sup>Since our data begins in 1999 and we need enough time period before the policy change on 2005, we will limit the rest of the analysis to look on treatment affects post-conception, which amounts to setting  $K = 0$  in Algorithm 1.

and finally compare them with the approaches currently used in the empirical literature.

## 4.1 Methodology

In this section, we formalize our policy analysis. Specifically, let  $W_{i,t}$  denote the level of the policy of interest unit  $i$  is exposed to in period  $t$ , and define  $W_i := (\dots, W_{i,T_{0,i}}, \dots, W_{i,T_{1,i}}, \dots)$  – the policy path unit for unit  $i$ . We assume that the policy path  $W_i$  is a known deterministic function of  $X_i$ ,  $W_i = W(X_i)$ , and thus is observed for every unit. For example, this assumption holds in applications where policies vary over locations as long as  $X_i$  includes the location of unit  $i$ , which is the case in our context.

For  $k \in \{-K, \dots, H\}$  we interpret the unobservable  $\tau_{i,e,k}$  defined in Assumption 2.1 as the realization of the underlying potential outcomes.

**Assumption 4.1. (POTENTIAL OUTCOMES)**

For every  $k \in \{-K, \dots, H\}$  and some known  $L_{k,a}, L_{k,b}$  we have:

$$\tau_{i,e,k} = \tau_{i,e,k}(\tilde{W}_{i,e,k}),$$

where  $\tilde{W}_{i,e,k} = (W_{i,e+k-L_{k,a}}, \dots, W_{i,e+k+L_{k,b}})$ .

$(L_{k,a}, L_{k,b})$  are user-specified parameters that describe which part of the policy path is relevant for the corresponding potential outcome. For example, if we set  $L_{k,a} = L_{k,b} \equiv 0$ , then  $\tilde{W}_{i,e,k} = W_{i,e+k}$ , thus measuring the exposure of unit  $i$  to the policy in period  $e+k$ . Assumption 4.1 then implies  $\tau_{i,e,k} = \tau_{i,e,k}(W_{i,e+k})$ , i.e., only contemporaneous level of policy exposure is relevant. By varying  $L_{k,a}$  and  $L_{k,b}$  Assumption 4.1 can accommodate various dynamic and anticipation effects. In our empirical application, we set  $L_{k,a} = k$ , and  $L_{k,b} = -k$ , which amounts to setting  $W_{i,e,k} = W_{i,e}$ , i.e., we assume that only the level of policy at the event time has a causal effect.

We denote  $\tilde{W}_{i,k} := \tilde{W}_{i,E_i,k}$ , and write:

$$\tau_{i,E_i,k} = \tau_{i,E_i,k}(\tilde{W}_{i,k}).$$

By construction,  $\tilde{W}_{i,k}$  is a deterministic function of  $X_i$  and  $E_i$ . As a result, it is impossible to distinguish the effect of policy from the effect of the event time  $e$  and heterogeneity in  $\tau_{i,e,k}$  across units. This situation mirrors the standard problem in policy evaluation, where without further assumptions, we cannot separate the effect of the introduction of a new policy from cross-temporal and cross-sectional heterogeneity in outcomes. To overcome this issue, we restrict the underlying heterogeneity and split  $\tau_{i,E_i,k}$  into three components.

**Assumption 4.2.** (*REGRESSION MODEL*)

For every  $k \in \{-K, \dots, H\}$  we have

$$\tau_{i,E_i,k} = \beta_k(X_i) + \mu_k(E_i) + \delta_{i,k}(\tilde{W}_{i,k}) + v_{i,k}, \quad \mathbb{E}[v_{i,k}|X_i, E_i, \tilde{W}_{i,k}] = 0.$$

The first two components of the regression function are analogous to conventional two-way fixed effects, which are defined over the two dimensions that create the variation in the treatment variable. The third part describes the effect of the policy, which potentially can vary over units. In Appendix B we discuss a causal model that implies Assumption 4.2.

Of course, we do not observe  $\tau_{i,E_i,k}$  but, as discussed in Section 2.4, we have a noisy estimate of it:

$$\hat{\tau}_{i,k} = \tau_{i,E_i,k} + \varepsilon_{i,k} + \xi_{i,k},$$

where  $\xi_{i,k}$  is the estimation error that vanishes as the sample size increases, while  $\varepsilon_{i,k}$  is the measurement error, which satisfies the orthogonality restrictions. Combining the pieces together, we can write

$$\hat{\tau}_{i,k} = \beta_k(X_i) + \mu_k(E_i) + \delta_{i,k}(\tilde{W}_{i,k}) + \tilde{v}_{i,k} + \xi_{i,k}, \quad \mathbb{E}[\tilde{v}_{i,k}|X_i, E_i, \tilde{W}_{i,k}, S_i^{tr} = 1] = 0, \quad (7)$$

where  $\tilde{v}_{i,k} = v_{i,k} + \varepsilon_{i,k}$ . We use this equation as the basis for the policy evaluation analysis.

**Remark 4.1.** Our policy analysis compares the values of  $\hat{\tau}_{i,k}$  across units with different values of  $X_i$ ,  $E_i$ , and  $\tilde{W}_{i,k}$ , and thus is cross-sectional in nature. However, by construction, the outcomes  $\hat{\tau}_{i,k}$  and policies  $\tilde{W}_{i,k}$  form a panel, and thus for each unit  $i$ , we can exploit the variation in  $k$ . A big advantage of having the panel dimension is the ability to relax Assumption 4.2 and allow for unit fixed effects. We do not do this because comparison over  $k$  is hard to interpret in the context of our empirical application.

**Remark 4.2.** Assumption 4.2 specifies a two-way structure that separates  $X_i$  and  $E_i$ . We opt for this choice because it is commonly used in applications. However, one can specify a different model that potentially includes interactions between  $X_i$  and  $E_i$ , as long as that structure still allows for residual variation in  $\tilde{W}_{i,k}$ . See Goldsmith-Pinkham et al. (2022) for a related discussion.

**Remark 4.3.** In our analysis, we treat unit-level data as i.i.d. which implies that errors  $v_{i,k}$  are i.i.d., and standard errors in estimating version of (7) can be constructed using the same weighted bootstrap scheme as we discussed in Section 2. In practice, one might want to treat data only as conditionally independent given some group-level variables and then conduct unconditional inference in (7) effectively allowing for correlation in  $v_{i,k}$ .

## 4.2 Institutional background – Dutch childcare provision expansion

### The 2005 Dutch childcare expansion reform

We begin with a brief overview of the subsidized childcare system and its recent changes in the Netherlands. Before the 2005 reform, the system for center-based daycare was funded differently across the board. Most daycare centers received subsidies from employers and local governments, which meant lower fees for parents compared to the 24% of daycare centers that weren't subsidized. However, these unsubsidized centers' costs could be partially reduced through tax deductions for working parents. In 2004, about 25% of children under the age of three were enrolled in such center-based care.

In addition to center-based care, around 25% of children also attended playgroups (peuterspeelzalen), which offered part-time care for less than four hours a day. This type of care was mostly used by families where one parent was not working. For children in primary school, there were both subsidized and unsubsidized options for out-of-school care, with costs for the unsubsidized care being partly tax-deductible. Yet, the enrolment rate for 4-12 year olds in centre-based care was less than 6% in 2004.

Before the reform, access to childcare was not uniform. Factors such as whether both parents' employers contributed to childcare costs, and varying policies between municipalities, affected accessibility and cost. There were also noticeable differences in what parents had to pay at childcare providers, regardless of whether companies or municipalities subsidized them.

The 2005 reform brought significant changes, creating a unified subsidy system for center-based care. From then on, all center-based daycare centers were eligible for the same government subsidy, which was given directly to parents using formal care. This change was especially beneficial for parents who used unsubsidized centers before 2005, as the new subsidy was typically more than what they saved through tax deductions.

### Childcare index

To create a municipality-based index of childcare supply, we utilized comprehensive information from the 5-digit sector classification, which has been available for all jobs in the Netherlands since 2001. This data allowed us to identify childcare workers and their job locations.<sup>7</sup> Our childcare supply index (*CCI*) for each municipality is calculated by dividing the number of childcare jobs in a given municipality  $m$  and year  $t$  ( $N_{m,t}^{jobs}$ ) by the number of children under

---

<sup>7</sup>Rabaté and Rellstab (2021) finds a strong correlation between the trends in childcare employment and the aggregate public spending on childcare, validating this index.

five years of age in the same locality ( $N_{m,t}^{children}$ ):

$$CCI_{m,t} = N_{m,t}^{jobs} / N_{m,t}^{children}$$

Figure IIIa presents the variation in childcare supply per preschool-aged children across municipalities, from 1999 to 2016. The dot represents the mean  $CCI$  in a given year where's the shaded area represent the distribution of  $CCI$  across municipalities in that year. The data reveal a significant range in the ratio of childcare workers to preschool children, with values ranging from 0 to 0.3.<sup>8</sup> The figure illustrates the substantial variation in childcare availability between different municipalities and the large increase due to the 2005 reform of childcare expansion.

## 4.3 Policy analysis

### Descriptive analysis

We begin by exploring the relationship between childcare provision levels and child-penalties (Figure A.2). We aggregate the estimated individual CP at the municipality times year-of-conception level, following the empirical strategy described in Section 3.2 and plot them against the childcare provision index ( $CCI$ ). The line represents the estimated coefficient from regressing the individual CP on  $CCI$ . We divide the estimation between men in blue and women in orange. We present the results for each estimate of the year relative to birth ( $k$ ). Figure A.2a presents the correlations using CP estimates for earnings. Similarly, Figure A.2b presents the correlations using CP estimates for participation.

We find that for women, childcare provision levels are related to lower child-penalty in participation and earnings. The relationship is stronger as we get farther from the timing of birth. This is consistent with the idea that childcare provision helps mothers return to the labor force after they finish their maternity leave, shrinking the CP. Fathers seem to present a similar relationship on the participation margin, although milder in strength.

A naive interpretation of these results will conclude that childcare provision reduces the CP. However, this interpretation will ignore the obvious supply-side response to a differential demand. In other words, households with higher employment and earnings potential will have higher willingness to pay for childcare services, resulting in higher availability. We, therefore, shift our analysis to explore the 2005 Dutch childcare expansion policy directly.

---

<sup>8</sup>It's important to note that some of the extreme values might be influenced by measurement errors, as the job location is tied to the firm's location and may not always align with the actual municipality where the job is performed.

## Semi-elasticity of childcare and labor supply

We shift our focus to illustrate how our flexible individual-level estimation allows us to estimate a common object of interest in economics – elasticities. In our analysis, we focus on  $CCI_{M_i, E_i}$ , which is the childcare index for the municipality  $M_i$  at the time  $E_i$  when the individual gave birth. We treat this variable as  $\tilde{W}_{i,k}$  from Assumption 4.1, which amounts to setting  $L_{k,a} = k$  and  $L_{k,b} = -k$ .

Specifically, we examine the semi-elasticity of childcare availability on labor supply by estimating a linear equation:

$$\frac{1}{6} \sum_{k=0}^5 \hat{\tau}_{i,k} = \delta \cdot CCI_{M_i, E_i} + \gamma_{M_i} + \beta_{A_i} + \mu_{E_i} + v_i \quad (8)$$

This equation is a special case of (7), aggregated over  $k$ , where we set  $\delta_{i,k}(\tilde{W}_{i,k}) = \gamma \tilde{W}_{i,k}$ . We control for the municipality fixed effects at the time of childbirth  $\gamma_{M_i}$ , year of giving birth fixed effects  $\mu_{E_i}$ , and the age-at-birth fixed effects  $\beta_{A_i}$ . By incorporating these fixed effects, the model aims to isolate the impact of the childcare expansion on the CP beyond potential confounding factors related to time, location, and age at childbirth.

We find a low and insignificant semi-elasticity of childcare provision and earnings (see Table I), consistent with other recent findings (Rabaté and Rellstab, 2021; Kleven et al., 2022).

## Nonlinearity

The results in Table I might be driven by the linear and constant relationship between CP and childcare provision imposed in Equation 8. Therefore, we exploit the flexibility of our measurement method and estimate non-parametric treatment effects:

$$\hat{\tau}_{i,k} = \delta_k(CCI_{M_i, E_i}) + \gamma_{k, M_i} + \beta_{k, A_i} + \mu_{k, E_i} + v_{i,k} \quad (9)$$

Compared to (8), this specification examines the relationship separately for each horizon  $k$ . It also allows for the effect of  $CCI_{M_i, E_i}$  to be nonlinear. In the context of Assumption 4.2 this amounts to specifying  $\delta_{i,k}(\tilde{W}_{i,k}) = \delta_k(\tilde{W}_{i,k})$ .

We split the estimation between men (blue) and women (orange) and present the results for each estimate of the year relative to birth ( $k$ ) in Figure IV. Each dot represents a binscatter of the non-parametric treatment effects estimated from equation 9. We can see in Figure IV a non-linear treatment effect of childcare provision expansion on almost all margins, groups, and years-post-birth. For example, the most striking non-linear effects are evident on the fathers' labor-force participation. These U-shaped effects suggest that an expansion of childcare has greater effects on father's labor supply when the baseline availability of childcare is low.

However, in higher levels of baseline childcare, a childcare expansion can also improve father's labor-force participation.

One could interpret these non-monotonic results as two different types of expansion implying different treatment effects. When childcare supply is low, the treatment provides more daycares, driving effects via the extensive margin. At higher levels, the treatment provides more after-hours care; consistent with the different margins of the policy that were enacted in 2007 for after-hours care expansion in schools. This non-linear result emphasizes the need for flexible estimation methods and can explain the wide range of results documented in the literature.

## Discussion

Our findings indicate a clear correlation between increased levels of childcare provision and reduced CP, highlighting the potential role of childcare availability in mitigating the labor market disadvantages faced by parents. However, it's notable that the elasticity of childcare provision in relation to earnings is relatively low and subject to change based on the specification of our model. This suggests a nuanced relationship where childcare provision does not uniformly translate to higher earnings for parents. Interestingly, we observe a non-linear effect on fathers' labor supply, indicating that the impact of childcare provision on labor participation is complex and varies by the type of expansion.

### 4.4 Comparison with common methods

In the last section, we discussed the benefits of employing individual event-study estimates as objects of interest for policy evaluation. We illustrated it in the context of assessing the effects of childcare expansion on the Child Penalty (CP). In this section, we will compare our method to common 'dynamic' policy-evaluation methods that tend to discretize their treatment effects and time variation, potentially overlooking substantial data variation. Our approach, in contrast, harnesses this variation to provide a more nuanced and detailed understanding of policy impacts. This not only enriches the analysis but also offers more transparent, precise, and actionable insights for policymakers and stakeholders. Nevertheless, we adjust in this section our flexible specification to a simplified binary specification so we can further compare the empirical results of the common one-step approach to our two-steps approach that splits the measurement and policy-evaluation steps.

## Common DiD specification for policy evaluation

The standard DiD method typically discretizes both the pre- and post-treatment periods, as well as the treatment. In this approach, the pre-treatment period includes individuals who gave birth in 2000, considering the Child Penalty (CP) for  $k = 0, \dots, 5$  as fully non-treated. The post-treatment period comprises individuals who gave birth in 2011, with CP for  $k = 0, \dots, 5$  deemed fully treated. Treatment is defined as municipalities with an expansion of at least 10 percentage points in  $CCI$  between pre and post-periods:

$$T_m \equiv \mathbf{1}\{\overline{CCI}_m^{2011-2016} - \overline{CCI}_m^{2000-2005} > 0.1\}.$$

This common DiD specification is illustrated in Figure IIIb. We can see that both treatment and control show parallel trends in the pre-period. In the post-period, the treatment diverges from the control group (mechanically). Importantly, the control group is also responding to the policy, which is expected given that the policy was implemented nationwide. This illustrates the downside of using this common approach. The key benefit of this approach is the ability to run it in one regression as follows:

$$\begin{aligned} Y_{i,t} = & \lambda_t + \alpha_{A_i} \\ & + \sum_{k \neq -1} \tau_k \mathbf{1}\{E_i + k = t\} + \sum_{k \neq -1} \beta_k \mathbf{1}\{E_i + k = t\} \mathbf{1}\{T_m\} \\ & + \sum_{k \neq -1} \gamma_k \mathbf{1}\{E_i + k = t\} \mathbf{1}\{E_i > 2005\} + \sum_{k \neq -1} \delta_k \mathbf{1}\{E_i + k = t\} \mathbf{1}\{E_i > 2005 \cdot T_m\} + \nu_{i,t} \quad (10) \end{aligned}$$

In contrast, our approach, designed to capture more nuanced treatment effects, involves two separate steps. First, we use Algorithm 1 to construct unit-level estimates  $\hat{\tau}_{i,k}$ . Second, we restrict the sample to births in 2000 and 2011 for individuals aged 30-34, aggregating  $\hat{\tau}_{i,k}$  by Treatment/Control group and Pre/Post period, and then computing the DiD estimator separately for each  $k$ .

Common intuition suggests that regression (10) should implement these steps automatically. However, the recent results for two-way methods (Sun and Abraham, 2021; Callaway and Sant'Anna, 2021) and regressions with multiple treatments (Goldsmith-Pinkham et al., 2022) show that this intuition is often misleading. By separating the analysis into two steps, we can guarantee that the measurement of  $\hat{\tau}_{i,k}$  is not affected by the particular implementation of the policy evaluation step.

## Comparing results

Aggregating individual-level estimates for different periods and locations we form four distinct panels (see Figure A.3). We then perform a DiD analysis on each dot in the panels, yielding the results depicted in the left panels of Figure V. Conversely, the common DiD specification, conceptually similar to ours and outlined in Equation 10, produces results illustrated in the right panels of Figure V.

The two approaches differ fundamentally in terms of sample restriction and accounting for unobserved heterogeneity. In our analysis, we include all individuals who had their first child from 2000 to 2022. We use this information to estimate  $\lambda_t(X_i)$  as precisely as possible. We then restrict the analysis to  $\hat{\tau}_{i,k}$  for individuals who had their first child in 2000 and 2011 at the age of 30-34. In contrast, the conventional approach (10) is based on the data from 2000 to 2005 and 2011 to 2016 and includes all individuals who had their first child at the age of 30-34. This makes the estimates of the CPs less precise. They also suffer from contamination bias (Goldsmith-Pinkham et al., 2022), though the magnitude of this bias is unclear.

The two approaches also differ substantially in terms of the underlying selection assumptions. Conventional specification implies that a simple cross-sectional comparison between individuals who had a child in a given year and those who would have it in the following year provides an unbiased estimator for the underlying CP. In contrast, Assumption 2.1 implies that such comparison can be confounded by individual-level fixed effects. Accounting for these effects has first-order implications for the results. In Figure A.4, we present the outcomes from our specification with and without unit fixed effects, and they are dramatically different.

These differences in sample selection result in substantial *qualitative* differences in the outcomes. For example, the results in Figure V suggest that a more extensive expansion of childcare provision could reduce the child penalty for mothers and fathers under our specification. Specifically, the left panels report the results from our specifications. We see that mothers and fathers reduce their labor supply slightly in the short term but increase it in the medium term (bottom panel). Similarly, the effect of childcare expansion on mothers' earnings is even greater than on fathers' earnings (top panel). These results are consistent with greater benefits of childcare expansion to mothers' earnings due to the combination of intra-household bargaining and social norms.

Conversely, the common specification yields the *opposite* conclusion. Under the common approach, both men and women *reduce* their labor supply and, therefore, their earnings due to a larger expansion in childcare provision. It is hard to think of the economic forces that would lead to a reduction in labor supply. Our discussion above suggests this result might be driven by an unaccounted unobserved heterogeneity and/or contamination bias.

**Remark 4.4.** The specification we use for comparison is motivated by the current status quo

in the empirical literature on CP (e.g., Kleven et al., 2019b). Of course, it is feasible to incorporate the unit fixed effects in the common specification (10) as well by dropping some of the indicators or allowing for never-treated individuals. However, doing this results in an additional form of contamination bias, as emphasized in (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Goldsmith-Pinkham et al., 2022). In particular, the coefficients in (10) correspond to weighted averages over multiple periods with negative weights, which can potentially result in misleading conclusions. In contrast, our approach does not suffer from this problem because it is based on aggregating  $\hat{\tau}_{i,k}$  for a pre-specified population of units.

### Beyond the child-penalty

The methodology used in this section demonstrates the benefits of event-study measurements as outcomes in a wider context. This flexible approach can be applied to examine the effects of various policies, providing a valuable tool for policy analysis and academic research. Importantly, by separating the analysis into two steps, we allow researchers to use all recent advances in policy evaluation techniques, including those with continuous and dynamic treatments. We view the flexibility and transparency of our approach as the main reason for using it in empirical practice.

## 5 Using individual estimates as a regressor (RHS)

So far, our analysis has been focused on using  $\hat{\tau}_{i,k}$  as a LHS variable in the regression-based analysis. As we have seen in the previous section, this econometric exercise can be useful for policy evaluation. In this section, we focus on using  $\hat{\tau}_{i,k}$  as a RHS variable. There are two problems that we need to address in this case. First, we need to explicitly account for the non-vanishing measurement error discussed in Section 2.4, which otherwise would lead to attenuation bias. Second, we might need to adjust for the correlation of the measurement error with the outcome we put on the LHS.

We start this section by developing a systematic way of addressing the first problem and discussing a way of solving the second one in a particular situation where the outcome of interest is  $\tau_{i,E_i,k}$  itself. As a byproduct of our analysis, we develop unit-specific measures of conditional variance, which we then investigate empirically. Finally, we apply our methodology to analyze the intergenerational transmission in CPs.

## 5.1 Methodology

In this section, we discuss the additional assumptions needed to adjust for the measurement error. Our identification argument and estimating strategy is based on applying and extending the approach from Arellano and Bonhomme (2012).

### Object of interest

To formalize our goal in this section, we first describe the object of interest. We focus on the coefficient  $\beta^*$  in the following partial linear model:

$$(\beta^*, f^*) := \arg \min_{\beta, f \in \mathcal{F}} \mathbb{E}[(Z_i - f(X_i, E_i) - \beta^\top \boldsymbol{\tau}_{i,E_i})^2 | S_i^{tr} = 1], \quad (11)$$

where  $\boldsymbol{\tau}_{i,E_i}^\top := (\tau_{i,E_k,-K}, \dots, \tau_{i,E_i,H})$  and  $\mathcal{F}$  includes all possible functions of  $(X_i, E_i)$ . Estimating (11) can be useful for different applications. For example, consider an intergenerational comparison, where variable  $(X_i, E_i, \boldsymbol{\tau}_{i,E_i})$  come from the generation of parents, while the outcome  $Z_i$  is recorded for a generation of children. Coefficient  $\beta^*$  then describes the correlation between  $Z_i$  and  $\boldsymbol{\tau}_{i,E_i}$  after adjusting for observed characteristics of parents. Indeed, the CP literature had been interested in the intergenerational transmission of CP as a measure of social norms stickiness across generations (Kleven et al., 2019b). More broadly, while the study of intergenerational mobility (IGM) has been focused on directly observed outcomes such as earnings, education, and wealth; a recent strand of the literature has attempted to measure the IGM in more abstract and self-constructed/imputed objects such as firm wage premium (Zohar and Dobbin, 2024; Engzell and Wilmers, 2024), skill (Cunha and Heckman, 2007), and financial literacy (Black et al., 2020).

By the standard Frisch-Waugh-Lovell Theorem (FWL Theorem), we can write the expression for  $\beta^*$ :

$$\beta^* = (\mathbb{E}[\mathbb{V}[\boldsymbol{\tau}_{i,E_i}|X_i, E_i]|S_i^{tr} = 1])^{-1} \mathbb{E}[(\boldsymbol{\tau}_{i,E_i} - \mathbb{E}[\boldsymbol{\tau}_{i,E_i}|X_i, E_i]) Z_i | S_i^{tr} = 1]$$

As a result, to identify  $\beta^*$  we need to identify the expectation of the conditional variance of  $\boldsymbol{\tau}_{i,E_i}$  and the expectation of the conditional covariance of  $\boldsymbol{\tau}_{i,E_i}$  and  $Z_i$  for the subpopulation of the treated units. We start by identifying the denominator, i.e., the expectation of the conditional variance:

$$\mathbb{E}[\mathbb{V}[\boldsymbol{\tau}_{i,E_i}|X_i, E_i]|S_i^{tr} = 1].$$

If we use  $\tau_{i,k}$  instead of  $\tau_{i,E_i,k}$  then the result is guaranteed to be biased:

$$\begin{aligned}\mathbb{V}[(\tau_{i,-H}, \dots, \tau_{i,K})^\top | X_i, E_i] &= \mathbb{V}[\boldsymbol{\tau}_{i,E_i} | X_i, E_i] + \mathbb{V}[\boldsymbol{\varepsilon}_i | X_i, E_i] + \\ &\quad \mathbb{E}[\boldsymbol{\varepsilon}_i \boldsymbol{\tau}_{i,E_i}^\top | X_i, E_i] + \mathbb{E}[\boldsymbol{\tau}_{i,E_i} \boldsymbol{\varepsilon}_i^\top | X_i, E_i],\end{aligned}$$

where  $\boldsymbol{\varepsilon}_i^\top = (\varepsilon_{i,-K}, \dots, \varepsilon_{i,H})$ . Assumption 2.2 guarantees that both cross-product terms are equal to zero,

$$\mathbb{E}[\boldsymbol{\tau}_{i,E_i} \boldsymbol{\varepsilon}_i^\top | X_i, E_i] = 0,$$

and thus, we have

$$\mathbb{V}[(\tau_{i,-H}, \dots, \tau_{i,K})^\top | X_i, E_i] = \mathbb{V}[\boldsymbol{\tau}_{i,E_i} | X_i, E_i] + \mathbb{V}[\boldsymbol{\varepsilon}_i | X_i, E_i]. \quad (12)$$

Since  $\tau_{i,k}$  is observed up to asymptotically vanishing estimation error  $\xi_{i,k}$  it follows that we can estimate the LHS of equation (12). To use this estimator for  $\mathbb{V}[\boldsymbol{\tau}_{i,E_i} | X_i, E_i]$  we need to adjust it for the bias term  $\mathbb{V}[\boldsymbol{\varepsilon}_i | X_i, E_i]$ .

## Identification and estimation

Without further restrictions, it is impossible to identify  $\mathbb{V}[\boldsymbol{\varepsilon}_i | X_i, E_i]$ . To solve this problem, we make another assumption.

**Assumption 5.1. (RESTRICTED HETEROSKEDASTICITY)**

Conditional variance of  $\boldsymbol{\epsilon}_i^\top = (\epsilon_{i,T_0,i}, \dots, \epsilon_{i,T_1,i})$  only varies with  $X_i$ ,

$$\mathbb{V}[\boldsymbol{\epsilon}_i | X_i, \boldsymbol{\tau}_i, \alpha_i] = \mathbb{V}[\boldsymbol{\epsilon}_i | X_i].$$

We use Assumption 5.1 to identify the variance matrix element-by-element. In particular, fix an event time  $e$ , and  $k, k' \in \{-K, \dots, H\}$ , where  $k$  can be potentially equal to  $e$ . Consider a unit  $i$  that is still far enough from its event  $E_i$ , such that  $S_i = 1$ ,  $E_i > e + \max\{k, k'\} + K$ , and  $e + \max\{k, k'\} + K \in \mathcal{T}_i^s$ . Using the notation introduced in Section 2 we can define:

$$\varepsilon_{i,e,k} := \Delta Y_{i,e,k} - \Delta \lambda_{e,k}(X_i) = \left( \epsilon_{i,e+k} - \frac{\sum_{T_0,i \leq l < e-K} \epsilon_{i,l}}{e - T_{0,i} - K} \right),$$

and analogously for  $k'$ . We use these errors to compute a covariance function:

$$\sigma_{e,k,k'}^\varepsilon(x) := \mathbb{E}[\varepsilon_{i,e,k} \varepsilon_{i,e,k'}, | X_i = x, S_i^{cont}(e + \max\{k, k'\} + K) = 1, e + \max\{k, k'\} + K \in \mathcal{T}_i^s].$$

**Algorithm 2:** Variance construction

**Data:**  $\mathcal{D}, K, H, k, k', \left\{\Delta\hat{\lambda}_{e,k}, \Delta\hat{\lambda}_{e,k'}\right\}_{e \in \mathcal{E}}, \{\hat{\tau}_{i,k}, \hat{\tau}_{i,k'}\}_{i \in S^{tr}} \mathcal{F}, \{\omega_i\}_{i \in \mathcal{S}}$

**Result:**  $\{\hat{\sigma}_{i,k,k'}\}_{i \in \mathcal{S}^{tr}}, \{\hat{\sigma}_{i,k,k'}^{obs}\}_{i \in \mathcal{S}^{tr}}$

```

1 for  $e \in \mathcal{E}$  do
2   Define
   
$$\hat{\tau}_{e,k} = \arg \min_{f \in \mathcal{F}} \sum_{i \in \mathcal{S}^{tr}(e)} (\hat{\tau}_{i,k} - f(X))^2 \omega_i,$$

   and analogously define  $\hat{\tau}_{e,k'}$ ;
3   for  $i : S_i^{cont}(e + \max\{k, k'\} + K), e + \max\{k, k'\} \in \mathcal{T}_i^s$  do
4     Define
     
$$\hat{\varepsilon}_{i,e,k} := \Delta Y_{i,e,k} - \Delta\hat{\lambda}_{e,k}(X_i),$$

     and analogously for  $k'$ ;
5   end
6   Define
   
$$\hat{\sigma}_{e,k,k'}^\varepsilon = \arg \min_{f \in \mathcal{F}} \sum_{\substack{i \in \mathcal{S}^{cont}(e + \max\{k, k'\} + K), \\ e + \max\{k, k'\} + K \in \mathcal{T}_i^s}} (\hat{\varepsilon}_{i,e,k} \hat{\varepsilon}_{i,e,k'} - f(X_i))^2 \omega_i$$

   for  $i \in \mathcal{S}^{tr}(e)$  do
7     Define  $\hat{\sigma}_{i,k,k'}^{obs} := (\hat{\tau}_{i,k} - \hat{\tau}_{e,k}(X_i))(\hat{\tau}_{i,k'} - \hat{\tau}_{e,k'}(X_i));$ 
8     Define  $\hat{\sigma}_{i,k,k'} := \hat{\sigma}_{i,k,k'}^{obs} - \hat{\sigma}_{e,k,k'}^\varepsilon(X_i);$ 
9   end
10 end

```

Assumption 2.2 and 5.1 then guarantee

$$(\mathbb{V}[\varepsilon_i | X_i, E_i])_{k,k'} = \sigma_{E_i, k, k'}^\varepsilon(X_i).$$

Repeating these steps for all values  $k, k' \in \{-K, \dots, H\}$  we identify the conditional variance matrix of  $\varepsilon_i$ . Using (12) we can thus identify the conditional variance matrix  $\mathbb{V}[\boldsymbol{\tau}_{i,E_i} | X_i, E_i]$ , and its expectation  $\mathbb{E}[\mathbb{V}[\boldsymbol{\tau}_{i,E_i} | X_i, E_i] | S_i^{tr} = 1]$ .

Algorithm 2 describes an implementation of this identification argument in practice. This algorithm uses  $\hat{\tau}_{i,k}$  and  $\hat{\tau}_{i,k'}$  delivered by Algorithm 1 as inputs. We rely on them to estimate the conditional mean functions  $\hat{\tau}_{e,k}(x)$  and  $\hat{\tau}_{e,k'}(x)$  (Step 2), and to construct unit-specific covariance measures  $\hat{\sigma}_{i,k,k'}$  (Step 7). The implementation of the latter step relies on estimates  $\hat{\sigma}_{e,k,k'}^\varepsilon(x)$  (Step 6), which is based on  $\hat{\varepsilon}_{i,e,k}$  and  $\hat{\varepsilon}_{i,e,k'}$ , which are empirical analogs of  $\varepsilon_{i,e,k}$  and  $\varepsilon_{i,e,k'}$  (Step 4). The algorithm delivers unit-specific covariance measures  $\hat{\sigma}_{i,k,k'}$  which we can average over all units in  $S^{tr}$  to estimate the corresponding element of  $\mathbb{E}[\mathbb{V}[\boldsymbol{\tau}_{i,E_i} | X_i, E_i] | S_i^{tr} = 1]$ .

We now turn to the numerator of  $\beta$ ,  $\mathbb{E}[(\boldsymbol{\tau}_{i,E_i} - \mathbb{E}[\boldsymbol{\tau}_{i,E_i}|X_i, E_i]) Z_i|S_i^{tr} = 1]$ . By definition

$$\tau_{i,k} - \tau_{E_i,k}(X_i) = \tau_{i,E_i,k} - \tau_{E_i,k}(X_i) + \varepsilon_{i,k},$$

and thus as long as  $\mathbb{E}[\varepsilon_{i,k} Z_i|S_i^{tr} = 1] = 0$ , we can identify the covariance:

$$\mathbb{E}[(\tau_{i,E_i,k} - \tau_{E_i,k}(X_i)) Z_i|S_i^{tr} = 1] = \mathbb{E}[(\tau_{i,k} - \tau_{E_i,k}(X_i)) Z_i|S_i^{tr} = 1].$$

The assumption that the measurement error  $\varepsilon_{i,k}$  is uncorrelated with  $Z_i$  is a substantive one and is not implied by any of the other assumptions we made in the paper. In practice, its validity depends on the nature of the outcome  $Z_i$ . We view this assumption as reasonable in the context of intergenerational comparisons, which amounts to assuming that the measurement error in the parents' generation is uncorrelated with the measurement error in the children's generation.

When this assumption is valid, then we can construct  $\hat{\beta}$  by first constructing the unnormalized coefficient that ignores the measurement error,

$$(\hat{\beta}_{un}, \hat{f}) := \arg \min_{\beta, f \in \mathcal{F}} \sum_{i \in \mathcal{S}^{tr}} (Z_i - f(X_i, E_i) - \beta^\top \hat{\tau}_{i,k})^2 \omega_i,$$

and then adjusting it:

$$\hat{\beta} := \hat{\Sigma}^{-1} \hat{\Sigma}^{obs} \hat{\beta}_{un},$$

where

$$(\hat{\Sigma})_{k,k'} := \frac{\sum_{i \in \mathcal{S}^{tr}} \hat{\sigma}_{i,k,k'} \omega_i}{\sum_{i \in \mathcal{S}^{tr}} \omega_i}, \quad (\hat{\Sigma}^{obs})_{k,k'} := \frac{\sum_{i \in \mathcal{S}^{tr}} \hat{\sigma}_{i,k,k'}^{obs} \omega_i}{\sum_{i \in \mathcal{S}^{tr}} \omega_i}.$$

## Discussion and validation

We now discuss how we can validate Assumption 5.1 and what happens if it fails. We start with the latter question, and for simplicity, ignore covariates  $X_i$ , assume that  $\mathcal{T}_i = \{0, 1, 2\}$ , and set  $K = 0$  and  $H = 1$ . In this case, the object of interest is  $\boldsymbol{\tau}_{i,1} = (\tau_{i,1,0}, \tau_{i,1,1})$ . For every unit  $i$  and  $k \in \{0, 1\}$  we compute

$$\Delta Y_{i,k,1} = Y_{i,1+k} - Y_{i,0}.$$

By definition, we have for  $i$  such that  $E_i = 1$ :

$$\tau_{i,k} = \Delta Y_{i,k,1} - \mathbb{E}[\Delta Y_{i,k,1}|E_i > 1 + k].$$

Similarly, we have for  $i$  such that  $E_i > 1 + k$

$$\varepsilon_{i,k,1} = \Delta Y_{i,k,1} - \mathbb{E}[\Delta Y_{i,k,1}|E_i > 1 + k].$$

For  $k, k' \in \{0, 1\}$  we compute the elements of the variance matrix using the following formula:

$$\mathbb{C}[\tau_{i,k}, \tau_{i,k'}|E_i = 1] - \mathbb{C}[\varepsilon_{i,k,1}, \varepsilon_{i,k',1}|E_i > \max\{k, k'\}].$$

Since constant shifts do not affect the covariance, we can equivalently write the previous expression as

$$\mathbb{C}[\Delta Y_{i,k,1}, \Delta Y_{i,k',1}|E_i = 1] - \mathbb{C}[\Delta Y_{i,k,1}, \Delta Y_{i,k',1}|E_i > \max\{k, k'\}]. \quad (13)$$

Now, suppose that  $E_i^*$  is assigned randomly, as it would be in a randomized experiment. Then, we can view expression (13) as a treatment effect on the appropriately transformed outcomes. In particular, if the event time has no impact on the underlying outcomes, then this difference is equal to zero. Motivated by this insight, we can consider an alternative comparison:

$$\mathbb{V}[\Delta Y_{i,0,1}|E_i = 2] - \mathbb{V}[\Delta Y_{i,0,1}|E_i > 2].$$

If  $E_i^*$  is randomly assigned and has no anticipation effects, then this difference should be equal to zero. It is instructive to compare this expression to the corresponding version for the means,

$$\mathbb{E}[\Delta Y_{i,0,1}|E_i = 2] - \mathbb{E}[\Delta Y_{i,0,1}|E_i > 2],$$

which corresponds to the standard comparison of average pretreatment outcomes. In empirical practice, this comparison is routinely used to test for “parallel” trends.

To implement this logic in practice, users simply need to execute Algorithms 1 and 2 with a larger value of  $K$ . For example, suppose that we believe that Assumption 2.1 is satisfied for  $K = 0$ , i.e., there are no anticipation effects. We then can run Algorithms 1 and 2 for  $K > 0$ , e.g.,  $K = 3$ . We then can use  $\hat{\tau}_{i,k}$  and  $\hat{\sigma}_{i,k,k'}$  for  $k, k' \in \{-K, \dots, -1\}$  for testing. In particular, comparing the average of  $\hat{\tau}_{i,k}$  to zero corresponds to the standard testing for the pretrends while comparing the average of  $\hat{\sigma}_{i,k,k'}$  to zero corresponds to the new test.

## 5.2 Estimating the variance in child-penalties

As we discussed in the last section, estimating the variance in CP has an *empirical value* for when one wants to use the unit-level estimates in the RHS. However, the variance in unit-level event studies might have an *economic value* by itself. For example, the variance in CP might

reflect the variability in labor supply post-childbirth, going beyond the means.

To investigate this, we apply Algorithm 2 to the same dataset we used in the previous two sections, with one additional sample restriction. The variance estimation is particularly sensitive to outliers, with a handful of extreme observations potentially changing the final result. To address this, we drop from the sample all individuals whose earnings were among the top 1% of their birth cohort at any time during their career. We focus on the variance components of the covariance matrix  $\hat{\sigma}_{i,k}^2 := \hat{\sigma}_{i,k,k}$ . First, we normalize  $\hat{\sigma}_{i,k}^2$ :

$$\tilde{\sigma}_{i,k}^2 := \frac{\hat{\sigma}_{i,k}^2}{\left( \frac{\sum_{j:E_j=E_i, G_j=G_i, B_j=B_i} \hat{Y}_{i,k}^c}{n_{g,b,e}} \right)^2}, \quad (14)$$

We use these coefficients to conduct the analysis similar to the one performed in Section 3. To convert these results to a meaningful scale, we take the square root of the resulting estimates, thus effectively computing the coefficient of variation (CV).

The results of this variance in CP for earnings present three interesting patterns (Figure VI). First, we can see the variance in earnings for me is increasing monotonically as we move further away from the event of giving birth (child's age). Importantly, while there is no difference between the variance three years pre-birth and the years before that (the baseline estimation), there is an increase in the variance starting already two years before the timing of birth. Second, while pre-birth the increase in the variance in CP is similar across men and women, they start diverging in the moment of birth, with women showing a decrease and stagnation in their earnings variance post-birth. This result is consistent with the naive k-means exercise in Figure Ib. Third, we can see that the measurement error correction works well for older parents (e.g., there is no differential variance up to three periods pre-birth), while this is not the case for younger parents in their late twenties. This is consistent with the results in Figure A.1a and serves as a diagnostic test for Assumption 5.1, suggesting again that we should focus on individuals who become parents in their early thirties.

We interpret these results in the following way. The wide heterogeneity in labor earnings pre-birth suggests potential diverging reasons for selection into parenthood: some choose the timing of parenthood since they lost/left a job and are ready to let their partner focus on their career while they raise the kid. Some possibly found a job and felt a greater sense of job security, leading them to opt into parenthood. These patterns are perhaps not gendered up until the later part of the career, when there are greater returns from intra-household specialization.

One side benefit of constructing unit-level measures of conditional variance and covariance is that we can use them to investigate the persistence in the CP. In particular, we consider

estimating the following regression coefficient

$$\beta_{0,5} = \frac{\mathbb{E}[\mathbb{E}[(\tau_{i,E_i,5} - \mathbb{E}[\tau_{i,E_i,5}|X_i, E_i])\tau_{i,E_i,0}|X_i, E_i]|S_i^{tr} = 1]}{\mathbb{E}[\mathbb{V}\tau_{i,E_i,0}|X_i, E_i]|S_i^{tr} = 1]},$$

This object is identified based on our previous result for the covariance matrix, and using the estimated covariance, a simple estimator of  $\beta$  is given by

$$\hat{\beta}_{0,5} := \frac{\sum_{i \in S^{tr}} \hat{\sigma}_{i,5,0}}{\sum_{i \in S^{tr}} \hat{\sigma}_{i,0,0}^2}.$$

We compute this estimator and find a coefficient around 1 for both men and women, which suggests that the individual CP are highly persistent.

### 5.3 Estimating the intergenerational mobility in child penalties

There are several relevant applications that illustrate how we can use unit-level estimates as regressors on the right-hand side (RHS). A natural application in the context of the CP is the intergenerational mobility (IGM) in CP, an object the literature had shown interest in as a measure of the importance of social norms in shaping the CP (Kleven et al., 2019b). However, to implement such an exercise empirically requires a long panel of individuals, labor-market, and fertility decisions that link parents and children across generations. Unfortunately, the Dutch data does not allow us to go far back enough to have meaningful labor-market outcomes for both parents and children's generations.

In order to illustrate how our method allows the CP-IGM estimation, we shift our analysis in this section to a longer panel dataset provided by the Panel Study of Income Dynamics (PSID). The PSID data structure encompasses annual data from 1968 to 1996 and biennial data from 1997 to 2021, allowing us to explore the evolution of CP across generations. Given the temporal structure of the dataset, the analysis is simplified into three steps:

#### Step 1: Separately estimate the parents and children's individual CP

We match parents and children using the Family Identification Mapping System (FIMS) and further limit our analysis to individuals with available outcome variables. For the parents' generation, our analysis spans the years 1968 to 1997, focusing on individuals who gave birth in 1970 or later. Similarly, we consider the children's generation, examining outcome variables from 1990 to 2021, with data collected biennially. We restrict this group to individuals born in 1992 or later. For both parents' and children's samples, we limit to individuals with childbirth occurring between the ages of 22 and 40 and trim individuals in the top 1% of earnings for

each year. Finally, we separately estimate CP using Algorithm 1 by gender and education. Resulting in the estimated individual CP for the children ( $\hat{\tau}_{i,k}$ ) and the parents ( $\hat{\tau}_{p(i),k}$ ).

### Step 2: Measurement error correction

One naive way to estimate the CP-IGM is to run the following regression between the CP of daughters and mothers averaged across all estimated years post-birth ( $k$ ):

$$\bar{\hat{\tau}}_i = \beta_0 + \beta^{IGM} \cdot \bar{\hat{\tau}}_{p(i)} + \epsilon_i$$

However, as we discussed extensively before, individual-level CP estimates suffer from measurement error, which creates an attenuation bias in  $\beta^{IGM}$  when used on the RHS. Specifically, we are concerned with the measurement in the parents' estimates. Therefore, we implement Algorithm 2 for the variance of individual-level CP within the parents' generation, similar to the procedure discussed in Section 5.2, resulting in the corrected variance  $\hat{\sigma}_{p(i)}^2$ .

### Step 3: Estimation of the intergenerational mobility coefficients

Finally, we manually calculate  $\beta^{IGM}$  so we can use the adjusted variance and thus correct the attenuation bias:

$$\beta^{IGM} := \frac{Cov(\bar{\hat{\tau}}_i, \bar{\hat{\tau}}_{p(i)})}{\hat{\sigma}_{p(i)}^2}$$

We compute the standard error for this coefficient using the weighted bootstrap discussed in Section 2.3.

## Results and discussion

We begin by plotting the scatter plot of daughters and mothers CP in Figure VII, using the estimates from Step 1. We focus on mothers and daughters since it is more reasonable for social norms to transmit from mothers to daughters rather than to sons (although these are feasible to estimate as well). The figure suggests some correlation, although it is hard to assess merely by staring at it. Therefore, we report in blue the results from the naive unadjusted coefficient (0.148), and in orange, the coefficient corrected for the measurement error, following the estimation procedure in Steps 2-3 above. We find a CP-IGM coefficient of 0.236, implying a 10% increase in mothers' CP is associated with a 2.36% higher CP for the daughters. This result suggests an economically important role of the transmission of social norms between generations.

Given the PSID's small sample size, it is important to take these specific estimates with a grain of salt. We believe that our estimates are very noisily estimated given the very small sample of merely 261 daughters and mothers pairs (due to our demanding sample restrictions). This, in turn, might imply that our variance correction is very crude. Therefore, we see this exercise as a proof of concept, which could be estimated in other CP studies using larger administrative data.

## 6 Conclusion

In this paper, we introduce a novel framework for analyzing unit-level effect heterogeneity in event studies, specifically applied to the study of child penalties (CP). Our method builds upon existing econometric techniques to offer a more nuanced understanding of how individuals are differently affected by events and policies, particularly in the context of childbirth and childcare provision.

By developing an estimation algorithm and employing a comprehensive dataset from the Netherlands, we document significant heterogeneity in CP and assess the impact of childcare policies on labor market outcomes. Our empirical exploration into the non-monotonic impacts of childcare provision policies on CP provides valuable insights into the complexity of policy effectiveness. We demonstrate that the success of such policies in mitigating CP is contingent on a multitude of factors, including baseline childcare levels and the specific dynamics of labor market participation among men and women. Moreover, our methodology enables the use of individual-level CP estimates as a regressor. This flexibility enriches the toolkit available to researchers, opening up new avenues for investigating the intergenerational transmission of CP and other user-constructed measures.

Our methodological and empirical contributions extend the scope of event study analysis by facilitating the direct measurement of unit-level effects and adjusting for biases associated with noisy estimates. The open-source tool 'unitdid' we provide enables researchers to apply our approach across various settings, enhancing the precision and depth of policy evaluation and economic research. Importantly, our findings underscore the necessity of individualized policy analysis, revealing how baseline conditions and specific individual circumstances significantly influence the effectiveness of interventions aimed at mitigating the CP.

By advocating for a shift towards more granular and informed analyses, this paper sets a new direction for economic research and policy-making. It reveals the limitations of one-size-fits-all policies and highlights the need for targeted interventions that consider the varied needs and circumstances of different population segments.

## References

- Andresen, Martin Eckhoff, and Emily Nix.** 2022a. “Can the Child Penalty Be Reduced? Evaluating Multiple Policy Interventions.”
- Andresen, Martin Eckhoff, and Emily Nix.** 2022b. “What Causes the Child Penalty? Evidence from Adopting and Same-Sex Couples.” *Journal of Labor Economics* 40 (4): 971–1004. [10.1086/718565](https://doi.org/10.1086/718565).
- Angelov, Nikolay, Per Johansson, and Erica Lindahl.** 2016. “Parenthood and the Gender Gap in Pay.” *Journal of Labor Economics* 34 (3): 545–579. [10.1086/684851](https://doi.org/10.1086/684851).
- Arellano, Manuel, and Stéphane Bonhomme.** 2012. “Identifying Distributional Characteristics in Random Coefficients Panel Data Models.” *The Review of Economic Studies* 79 (3): 987–1020. [10.1093/restud/rdr045](https://doi.org/10.1093/restud/rdr045).
- Arkhangelsky, Dmitry, and Guido Imbens.** 2023. “Causal Models for Longitudinal and Panel Data: A Survey.” November. [10.48550/arXiv.2311.15458](https://arxiv.org/abs/2311.15458).
- Athey, Susan, and Stefan Wager.** 2021. “Policy Learning With Observational Data.” *Econometrica* 89 (1): 133–161. [10.3982/ECTA15732](https://doi.org/10.3982/ECTA15732).
- Black, Sandra E., Paul J. Devereux, Petter Lundborg, and Kaveh Majlesi.** 2020. “Poor Little Rich Kids? The Role of Nature versus Nurture in Wealth and Other Economic Outcomes and Behaviours.” *The Review of Economic Studies* 87 (4): 1683–1725. [10.1093/restud/rdz038](https://doi.org/10.1093/restud/rdz038).
- Bonhomme, Stéphane, and Ulrich Sauder.** 2011. “Recovering Distributions in Difference-in-Differences Models: A Comparison of Selective and Comprehensive Schooling.” *The Review of Economics and Statistics* 93 (2): 479–494. [10.1162/REST\\_a\\_00164](https://doi.org/10.1162/REST_a_00164).
- Borusyak, Kirill, Xavier Jaravel, and Jann Spiess.** 2023. “Revisiting Event Study Designs: Robust and Efficient Estimation.” September.
- Callaway, Brantly, and Pedro H.C. Sant’Anna.** 2021. “Difference-in-Differences with Multiple Time Periods.” *Journal of Econometrics* 225 (2): 200–230. [10.1016/j.jeconom.2020.12.001](https://doi.org/10.1016/j.jeconom.2020.12.001).
- Castellanos, María Alexandra.** 2024. “Immigration, Parenthood and Child Penalties.”
- Chamberlain, Gary.** 1992. “Efficiency Bounds for Semiparametric Regression.” *Econometrica* 60 (3): 567–596. [10.2307/2951584](https://doi.org/10.2307/2951584).
- Chamberlain, Gary, and Guido W. Imbens.** 2003. “Nonparametric Applications of Bayesian Inference.” *Journal of Business & Economic Statistics*. [10.1198/073500102288618711](https://doi.org/10.1198/073500102288618711).

- Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val.** 2018. “Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India.” Technical report, National Bureau of Economic Research.
- Chernozhukov, Victor, Whitney K Newey, and Rahul Singh.** 2023. “A simple and general debiased machine learning theorem with finite-sample guarantees.” *Biometrika* 110 (1): 257–264.
- Cunha, Flavio, and James Heckman.** 2007. “The Technology of Skill Formation.” *American Economic Review* 97 (2): 31–47. [10.1257/aer.97.2.31](https://doi.org/10.1257/aer.97.2.31).
- Engzell, Per, and Nathan Wilmers.** 2024. “Firms and the Intergenerational Transmission of Labor Market Advantage.” preprint, SocArXiv. [10.31235/osf.io/mv3e9](https://doi.org/10.31235/osf.io/mv3e9).
- Foster, Dylan J., and Vasilis Syrgkanis.** 2023. “Orthogonal Statistical Learning.” *The Annals of Statistics* 51 (3): 879–908. [10.1214/23-AOS2258](https://doi.org/10.1214/23-AOS2258).
- Gallen, Yana.** 2019. “The Effect of Parental Leave Extensions on Firms and Coworkers.”
- Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár.** 2022. “Contamination Bias in Linear Regressions.” June. [10.3386/w30108](https://doi.org/10.3386/w30108).
- Karademir, Sencer, Jean-William P. Laliberté, and Stefan Staubli.** 2024. “The Multigenerational Impact of Children and Childcare Policies.” March. [10.3386/w32204](https://doi.org/10.3386/w32204).
- Kennedy, Edward H.** 2023. “Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects.” *Electronic Journal of Statistics* 17 (2): 3008–3049. [10.1214/23-EJS2157](https://doi.org/10.1214/23-EJS2157).
- Kleven, Henrik, Camille Landais, Johanna Posch, Andreas Steinhauer, and Josef Zweimüller.** 2019a. “Child Penalties across Countries: Evidence and Explanations.” *AEA Papers and Proceedings* 109 122–126. [10.1257/pandp.20191078](https://doi.org/10.1257/pandp.20191078).
- Kleven, Henrik, Camille Landais, Johanna Posch, Andreas Steinhauer, and Josef Zweimüller.** 2022. “Do Family Policies Reduce Gender Inequality? Evidence from 60 Years of Policy Experimentation.” August.
- Kleven, Henrik, Camille Landais, and Jakob Egholt Søgaard.** 2019b. “Children and Gender Inequality: Evidence from Denmark.” *American Economic Journal: Applied Economics* 11 (4): 181–209. [10.1257/app.20180010](https://doi.org/10.1257/app.20180010).
- Lim, Nayeon, and Lisa-Marie Duletzki.** 2023. “The Effects of Public Childcare Expansion on Child Penalties - Evidence From West Germany.”

- Lloyd, S.** 1982. “Least Squares Quantization in PCM.” *IEEE Transactions on Information Theory* 28 (2): 129–137. [10.1109/TIT.1982.1056489](https://doi.org/10.1109/TIT.1982.1056489).
- Nie, X, and S Wager.** 2021. “Quasi-Oracle Estimation of Heterogeneous Treatment Effects.” *Biometrika* 108 (2): 299–319. [10.1093/biomet/asaa076](https://doi.org/10.1093/biomet/asaa076).
- Rabaté, Simon, and Sara Rellstab.** 2021. “The Child Penalty in the Netherlands and Its Determinants.” *CPB Discussion Paper*. [10.34932/TRKZ-QH66](https://doi.org/10.34932/TRKZ-QH66).
- Robinson, P. M.** 1988. “Root-N-Consistent Semiparametric Regression.” *Econometrica* 56 (4): 931–954. [10.2307/1912705](https://doi.org/10.2307/1912705).
- Semenova, Vira, and Victor Chernozhukov.** 2021. “Debiased machine learning of conditional average treatment effects and other causal functions.” *The Econometrics Journal* 24 (2): 264–289.
- Sun, Liyang, and Sarah Abraham.** 2021. “Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects.” *Journal of Econometrics* 225 (2): 175–199. [10.1016/j.jeconom.2020.09.006](https://doi.org/10.1016/j.jeconom.2020.09.006).
- Zohar, Tom, and Caue Dobbin.** 2024. “Quantifying the Role of Firms in Intergenerational Mobility.”

# Tables

Table I: Low Semi-Elasticity of Childcare Provision and Earnings

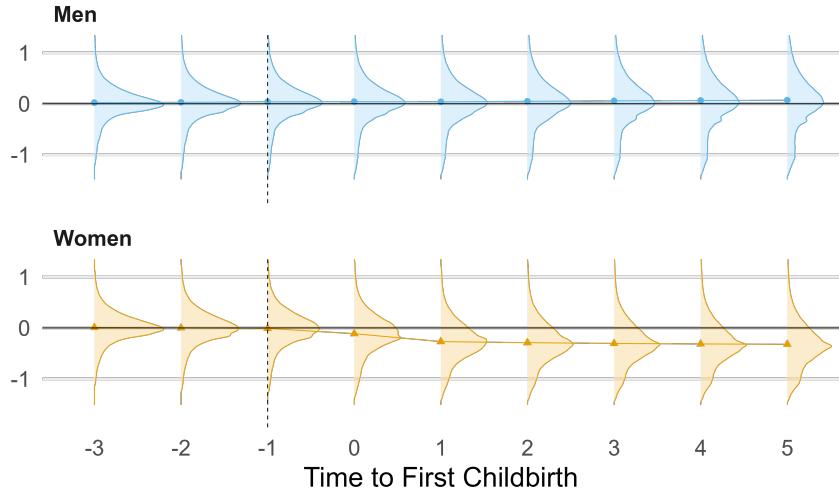
	Men		Women	
	CP in Earnings	CP in Participation	CP in Earnings	CP in Participation
Childcare Index	0.091 (0.073)	-0.036 (0.040)	0.066 (0.038)	0.014 (0.018)
Observations	263904	263904	262490	262490
R <sup>2</sup>	0.01	0.00	0.05	0.01
FE: Year at Birth	X	X	X	X
FE: Age at Birth	X	X	X	X
FE: Municipality	X	X	X	X

This table presents the elasticity of child-penalty to childcare provision following Equation 8. Each regression includes fixed effects for the year of birth, age at birth, and municipality. Standard errors are clustered by year and municipality at birth. \*\* p < .01; \* p < .05; +p < .1.

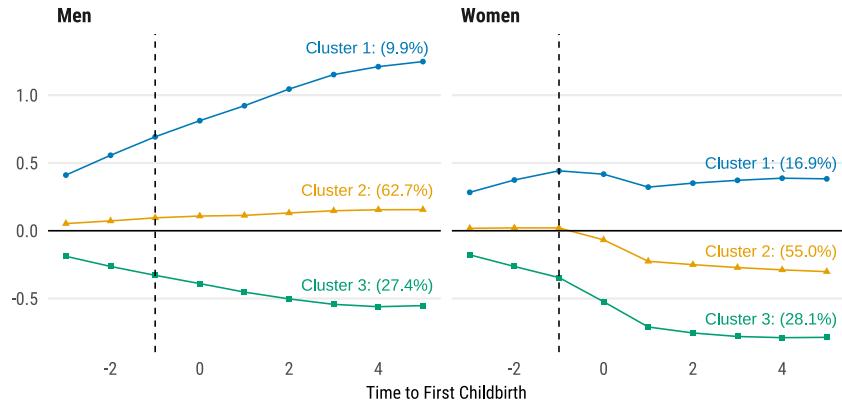
# Figures

Figure I: High Heterogeneity in Child-penalties (CP)

(a) Distribution of Individual CP

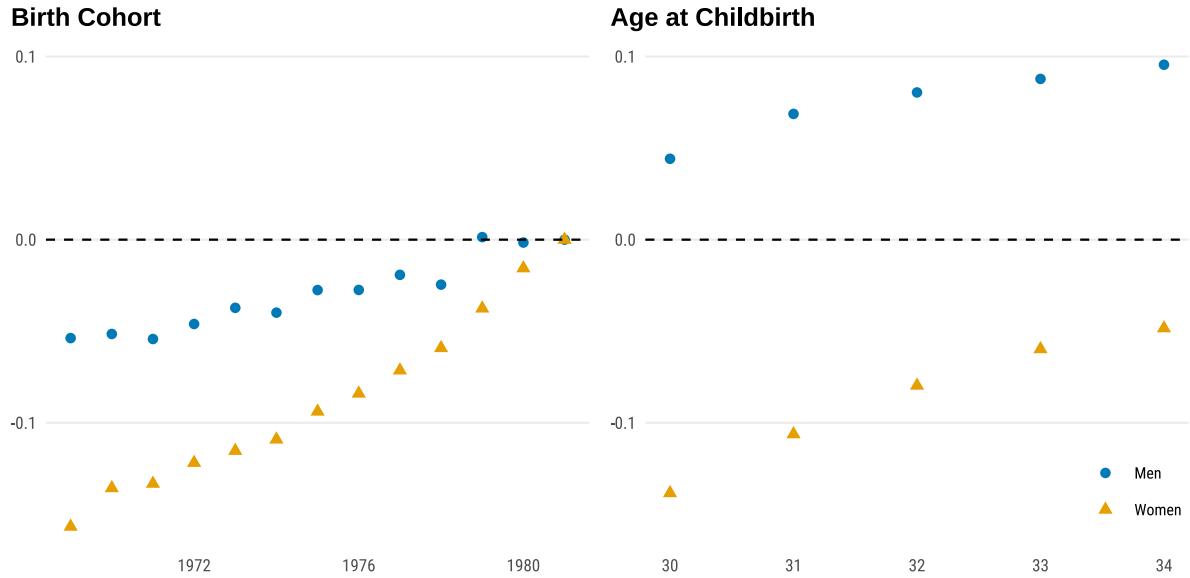


(b) Different CP Paths (K-Means)



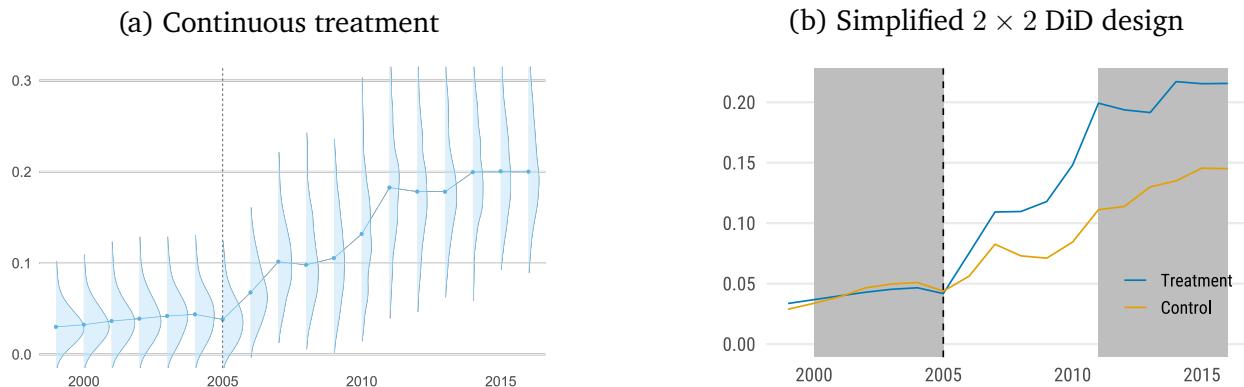
*Notes:* These figures presents the variation in child-penalty estimates (CP). Figure Ia plots the marginal distributions of  $\tilde{\tau}_{i,k}$  from Equation 6 for earnings, by time relative to first childbirth and genders, pooled across all birth cohorts. The dots represent the mean of each distribution. Figure Ib shows the results from applying a K-means algorithm to the vector of estimated child penalties  $\tilde{\tau}_i$  and classify all individuals of the same gender into three groups.

Figure II: Avg. CP Decomposition by Birth Cohort and Age at First Childbirth



Notes: This figure presents the CP estimates aggregated and decomposed by observables of interest – age of first childbirth and birth cohort, as described in Section 3.2. All estimates are normalized to the CP of the 1980 birth cohort.

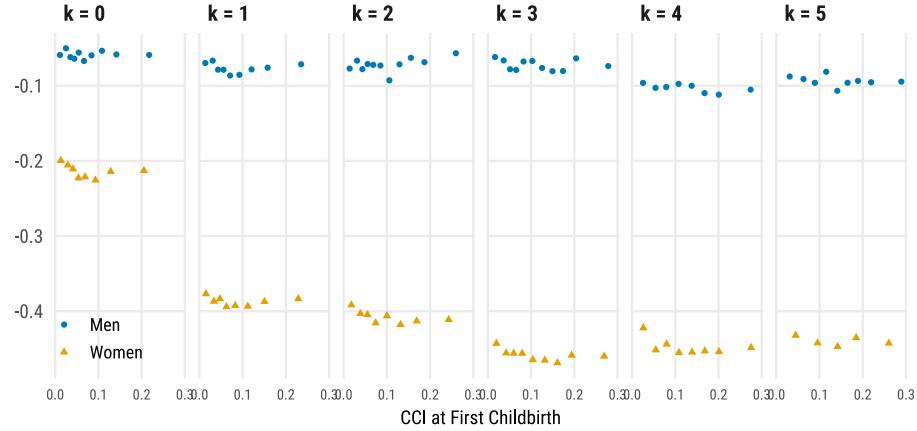
Figure III: Continuous and discrete variation of childcare supply expansion



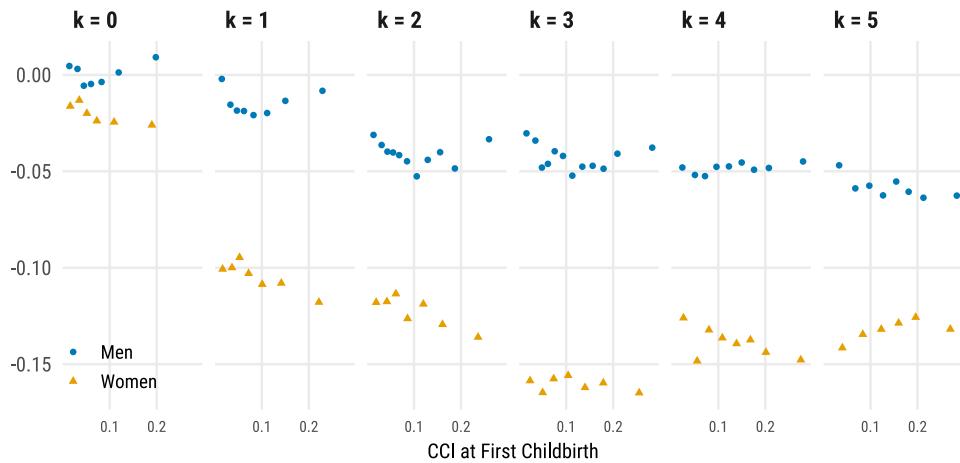
Notes: These figures presents the variation in childcare supply per preschool-aged children across municipalities, from 1999 to 2016. Our childcare supply index ( $CCI$ ) for each municipality is calculated by dividing the number of childcare jobs in a given municipality  $m$  and year  $t$  ( $N_{m,t}^{jobs}$ ) by the number of children under five years of age in the same locality ( $N_{m,t}^{children}$ ). The vertical line illustrates the timing of the 2005 Dutch childcare expansion reform. Panel (a) illustrates the substantial variation in childcare availability between different municipalities and the large increase due to the 2005 reform of childcare expansion. The dot represents the mean  $CCI$  in a given year where's the shaded area represent the distribution of  $CCI$  across municipalities in that year. Panel (b) illustrates the equivalent simplified  $2 \times 2$  DiD design, where the time variation is binary (grey area) and treatment is binary (see Section 4.4). The pre-treatment period includes individuals who gave birth in 2000, considering the Child Penalty (CP) for  $k = 0, \dots, 5$  as fully non-treated. The post-treatment period comprises individuals who gave birth in 2011, with CP for  $k = 0, \dots, 5$  deemed fully treated. Treatment is defined as municipalities with an expansion of at least 10 percentage points in  $CCI$  between pre and post periods.

Figure IV: Non-parametric treatment-effects of childcare provision expansion on child-penalties

(a) Earnings

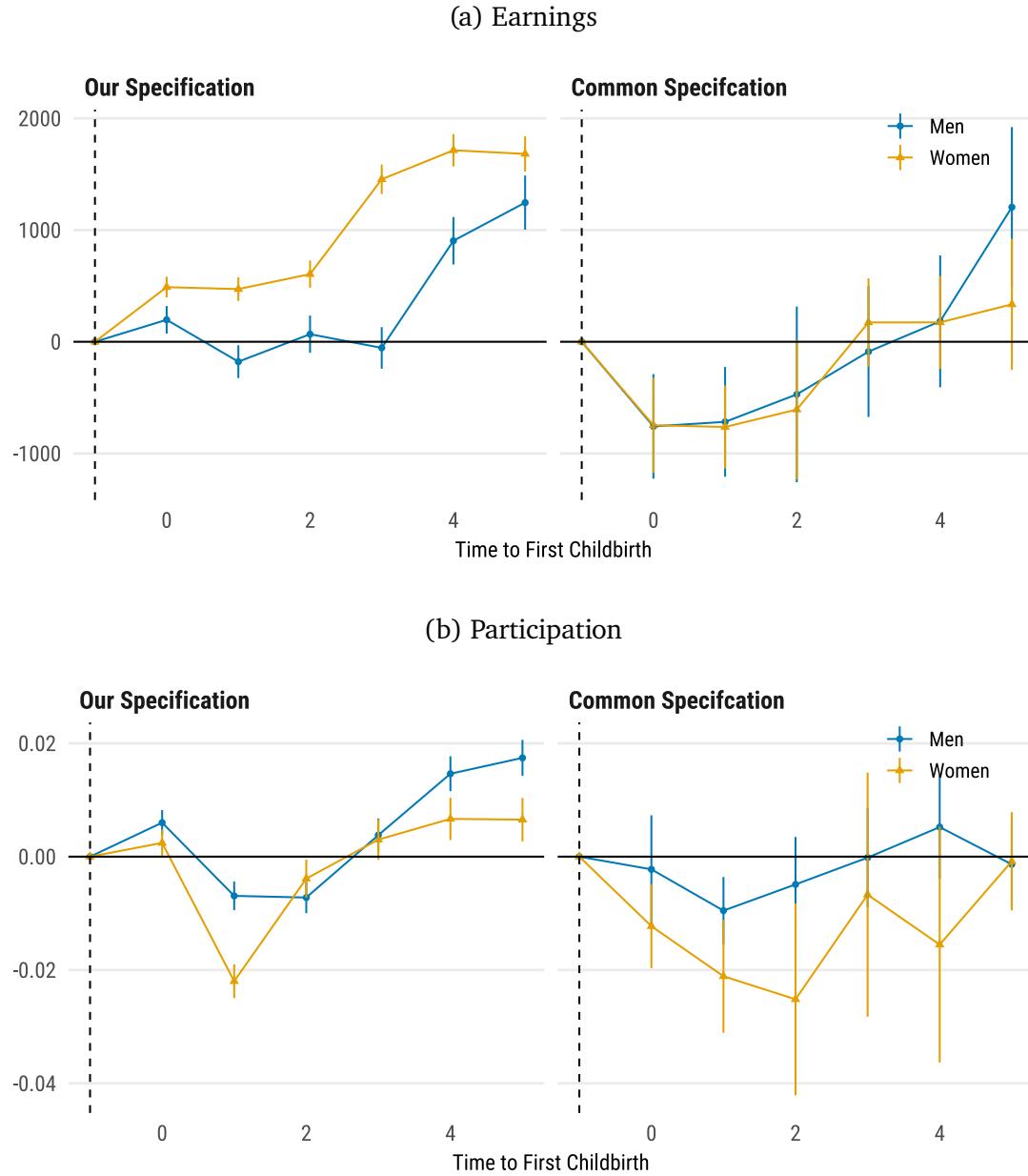


(b) Participation



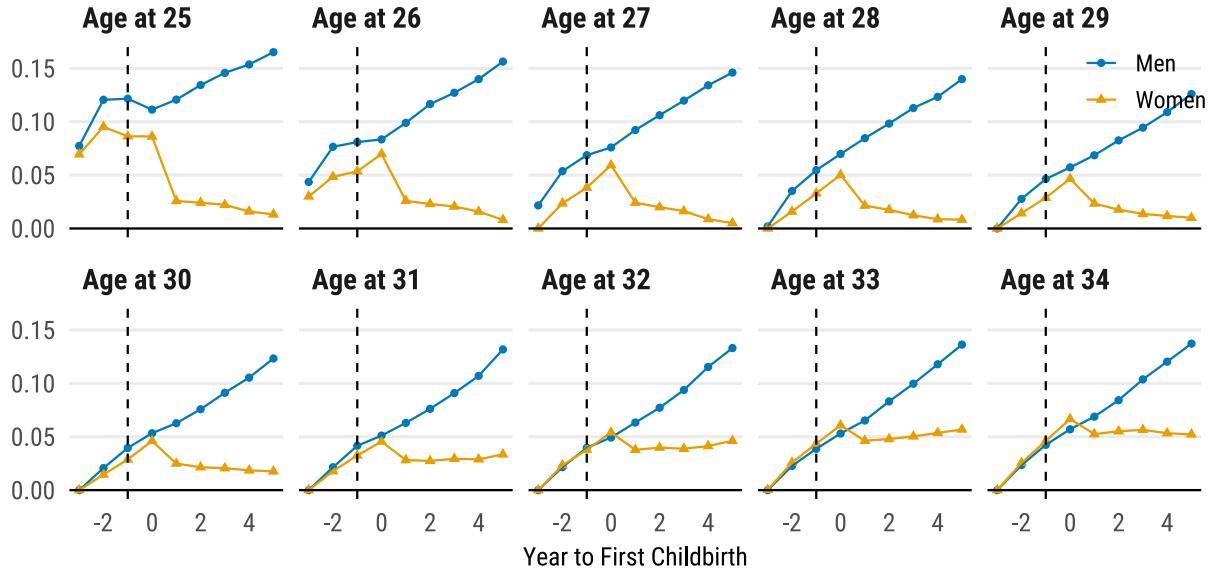
*Notes:* This figure presents the effect of childcare provision expansion on child-penalties (CP). We split the estimation between men (blue) and women (orange). Each dot represents a binscatter of the non-parametric treatment effects estimated from equation 9. We present the results for each estimate of the year relative to birth ( $k$ ). We estimate individual-CP following the empirical strategy described in Algorithm 1. Our childcare supply index ( $CCI$ ) for each municipality is calculated by dividing the number of childcare jobs in a given municipality  $m$  and year  $t$  ( $N_{m,t}^{jobs}$ ) by the number of children under five years of age in the same locality ( $N_{m,t}^{children}$ ).

Figure V: Large qualitative differences between ours' and the common policy evaluation method



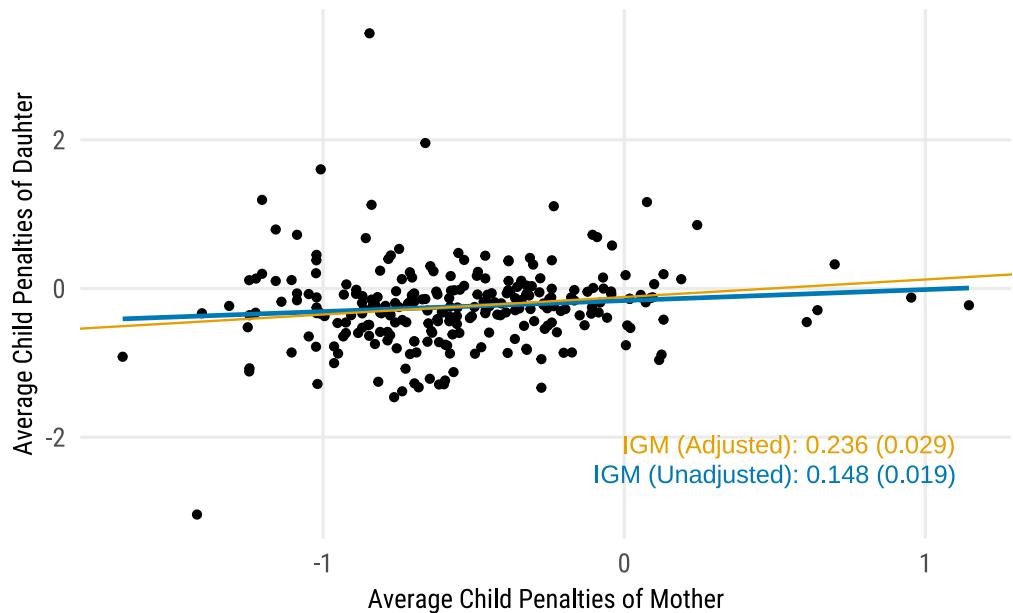
Notes: This figure presents the discrete diff-in-diff estimates of the effects of childcare expansion on the child-penalty using different estimation strategies. Panel (a) presents the results for earnings, while Panel (b) presents the results for participation. The aggregation process in Step 2 of our approach generates four distinct panels (see Figure A.3). We then perform a diff-in-diff analysis on each of the corresponding dots in the panels of Figure A.3. Conversely, the common DiD specification, conceptually similar to ours and outlined in Equation 10, produces results illustrated in the right panels.

Figure VI: Distributional properties – Variance in child-penalties



*Notes:* This figure presents the impact of childbirth on the earnings variability among different parents, split by gender and age of giving birth. We adopt the steps in Algorithm 2 for the case of the child-penalty to implement this estimation, as described in Section 5.2.

Figure VII: Intergenerational mobility in CP: mothers vs daughters

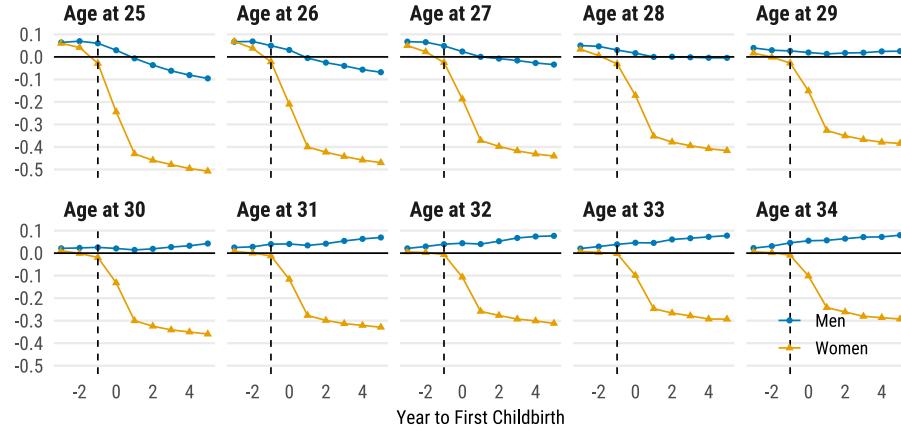


*Notes:* This figure presents a scatter plot and the correlation between child-penalties (CP) across generations. We take the average individual-CP across five years post birth for mothers and daughters. We follow the empirical strategy described in Section 5.3. The blue line represents the estimated coefficient from regressing the individual-CP of the daughters on their mothers' CP, adjusting for ME. The orange line represent the corresponding coefficient, adjusted for the ME based on the estimation procedure in Steps 2-3 of Section 5.3.

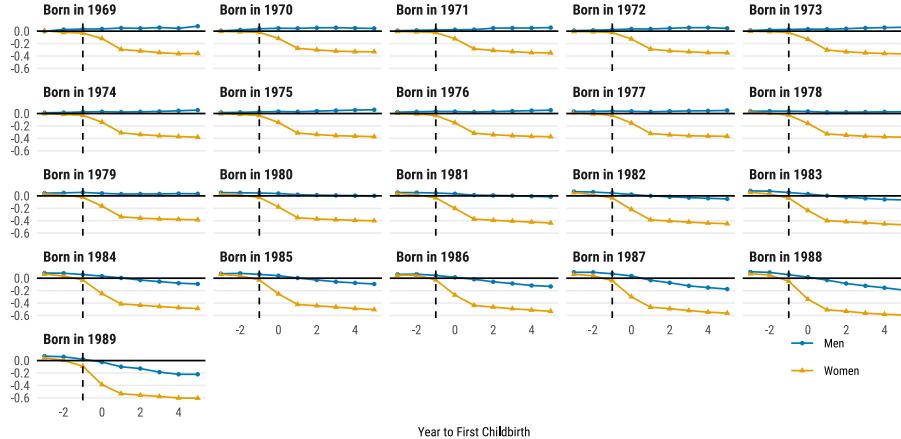
## A Figures and tables

Figure A.1: Avg. CP by Birth Cohort and Age at First Childbirth

(a) Aggregation by Age at First Childbirth



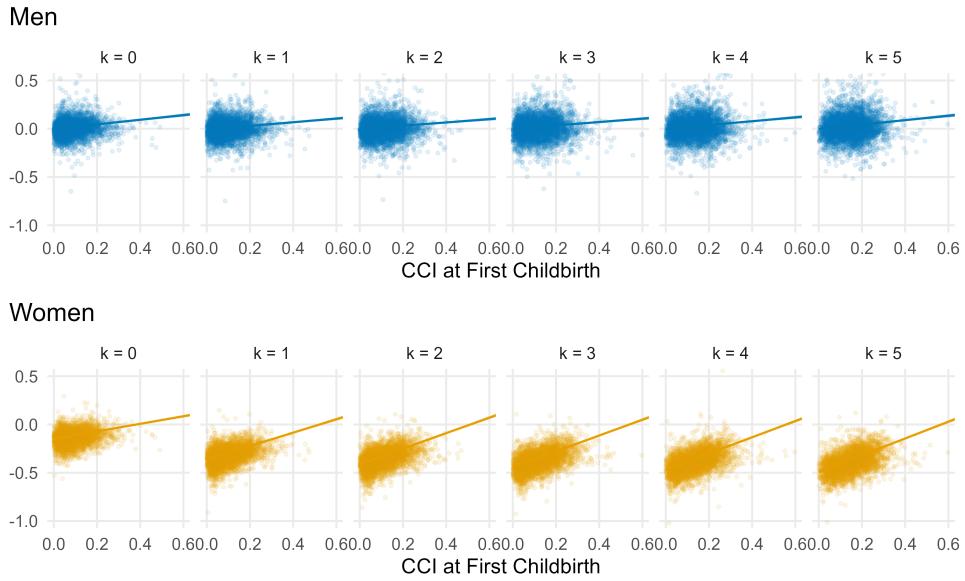
(b) Aggregation by Birth Cohort



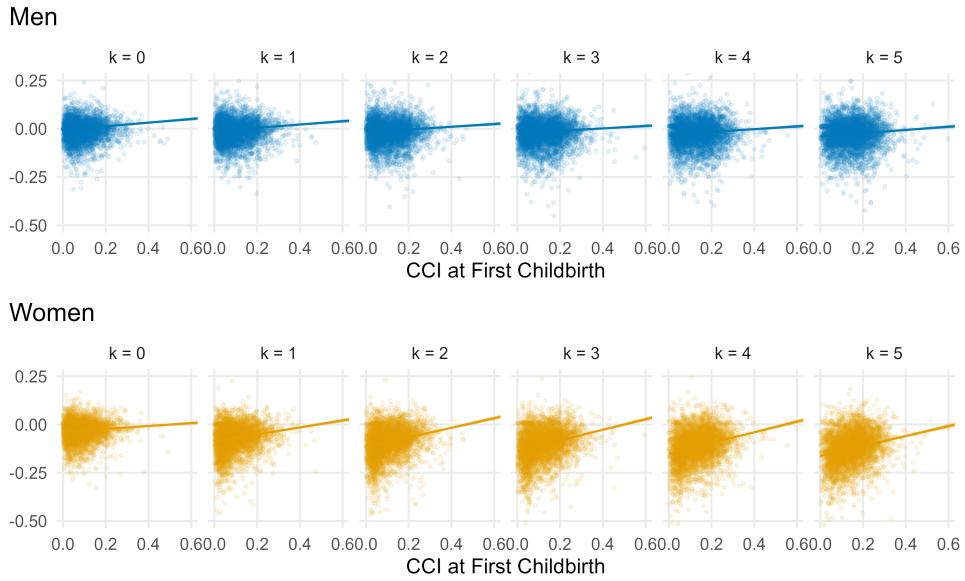
*Notes:* This figure presents the CP estimates aggregated by observables of interest – age of first childbirth and birth cohort, as described in Section 3.2. Panel (a) reports the average CP across age at first childbirth at all horizons. Similarly, Panel (b) reports the average CP across birth cohorts at all horizons post-birth.

Figure A.2: Childcare provision levels are correlated with lower child-penalty

(a) Earnings



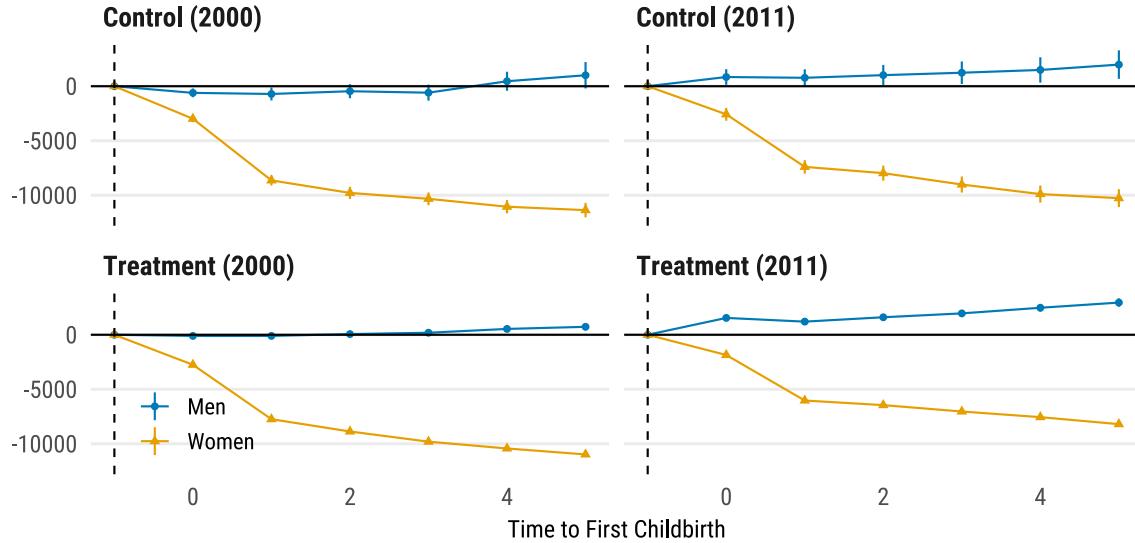
(b) Participation



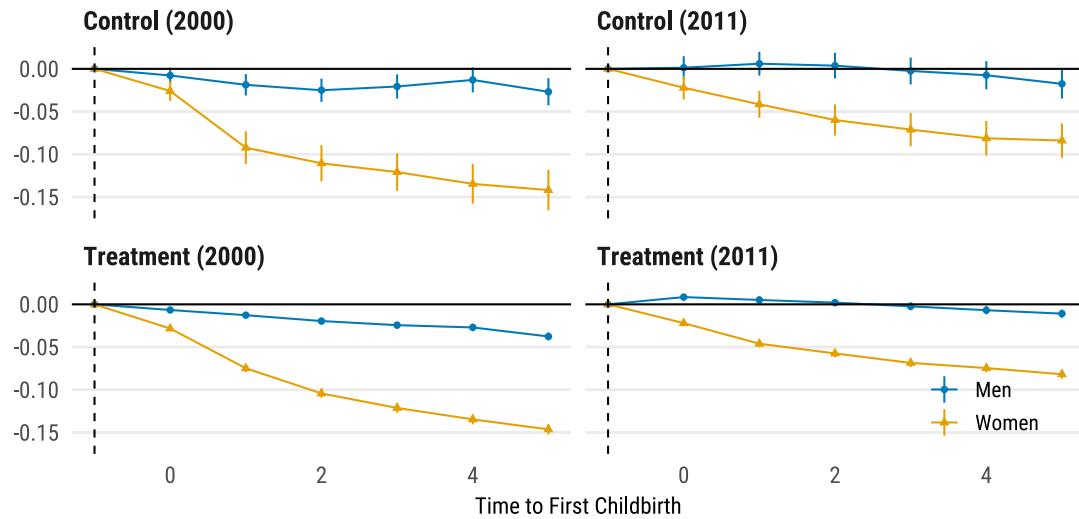
*Notes:* This figure presents the correlation between child-penalties (CP) and childcare provision index ( $CCI$ ). We aggregate the estimated individual-CP at the municipality times year-of-conception level, following the empirical strategy described in Section 3.2 and plot them against the childcare provision index ( $CCI$ ). The line represents the estimated coefficient from regressing the individual-CP on  $CCI$ . We divide the estimation between men in blue and women in orange. We present the results for each estimate of the year relative to birth ( $k$ ). Figure A.2a presents the correlations using CP estimates for earnings. Similarly, Figure A.2b presents the correlations using CP estimates for participation. Our childcare supply index ( $CCI$ ) for each municipality is calculated by dividing the number of childcare jobs in a given municipality  $m$  and year  $t$  ( $N_{m,t}^{jobs}$ ) by the number of children under five years of age in the same locality ( $N_{m,t}^{children}$ ).

Figure A.3: Child-penalty by treatment and post status

(a) Earnings



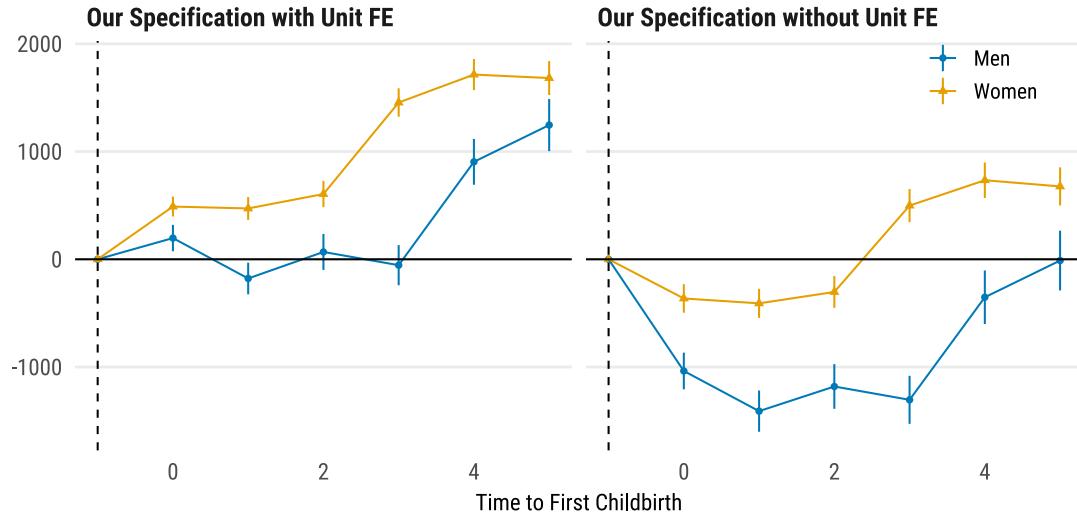
(b) Participation



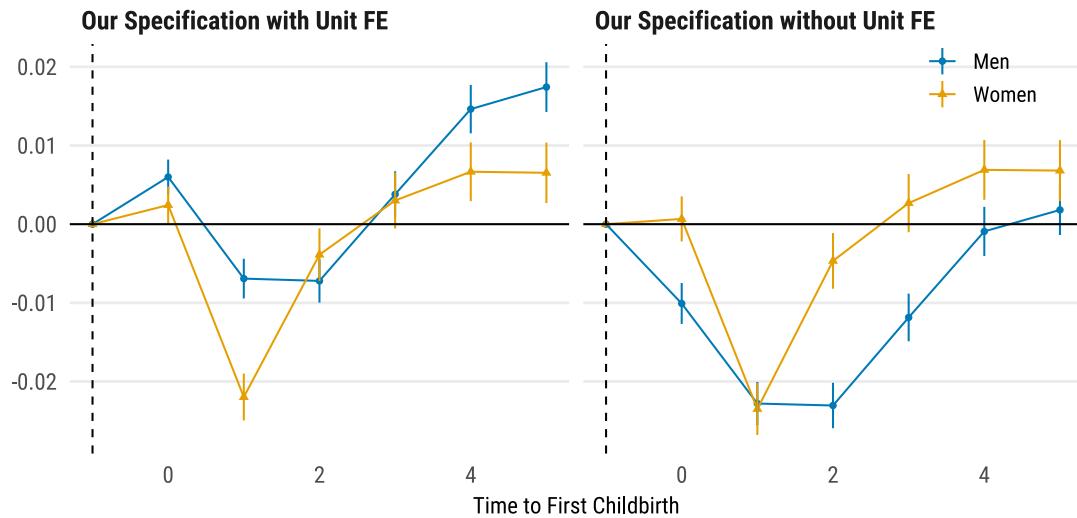
*Notes:* This figure presents the child-penalty estimates split by treatment status, and pre- and post- the childcare policy expansion. Panel A.3a presents the results for earnings, while Panel A.3b presents the results for participation. The pre-treatment period includes individuals who gave birth in 2000, considering the Child Penalty (CP) for  $k = 0, \dots, 5$  as fully non-treated. The post-treatment period comprises individuals who gave birth in 2011, with CP for  $k = 0, \dots, 5$  deemed fully treated. Treatment is defined as municipalities with an expansion of at least 10 percentage points in CCI between pre and post periods (see Figure IIIb and Section 4.4).

Figure A.4: Large differences due to downward selection bias (Unit FE)

(a) Earnings



(b) Participation



Notes: This figure presents the discrete diff-in-diff estimates of the effects of childcare expansion on the child penalty using different estimation strategies, illustrating the sensitivity to unit fixed effect. Panel (a) presents the results for earnings, while Panel (b) presents the results for participation. Both panels The aggregation process in Step 2 of our approach generates four distinct panels (see Figure A.3). We then perform a diff-in-diff analysis on each of the corresponding dots in the panels of Figure A.3. The left panels show our results with unit fixed effect, while the right panels show the results without.

## B Additional theoretical derivations

### Causal model for policy analysis

In this section we sketch a causal model behind Assumption 4.2. We start by assuming that the potential outcomes can be separated into two parts:

$$\tau_{i,e,k}(w) = \alpha_{i,k}(e) + \gamma_{i,k}(w). \quad (\text{B.1})$$

This restriction eliminates state dependence: the policy's effect is the same regardless of the underlying event time. Note that it still allows for unrestricted heterogeneity in effects, so in the sample, the policy's effect at realized event times can be different. In other words, we resolve the standard trade-off between state dependence and unobserved heterogeneity in favor of the latter.

Next, we split  $\alpha_{i,k}(e)$  into the systematic and residual part:

$$\alpha_{i,k}(e) = \alpha_k(X_i, E_i, e) + \nu_{i,k}(e), \quad \mathbb{E}[\nu_{i,k}(e)|X_i, E_i] = 0.$$

Note that this separation is mechanical and does not impose any restrictions on the correlation between  $E_i$  and  $\alpha_{i,k}(e)$ . As a result, it is not entirely standard, because the term  $\alpha_k(X_i, E_i, e)$  depends on both observed  $E_i$  and potential  $e$ . If we were to assume that  $E_i$  is as good as randomly assigned given  $X_i$ , then  $\alpha_{i,k}(X_i, E_i, e)$  would have depended solely on  $e$ .

Next, we make a functional form assumption and evaluate the resulting components at  $E_i$ :

$$\alpha(X_i, E_i, e) = \beta_k(X_i) + \mu_k(E_i, e) \Rightarrow \alpha(X_i, E_i, E_i) = \beta_k(X_i) + \mu_k(E_i, E_i). \quad (\text{B.2})$$

This assumption imposes a restriction on the underlying effect heterogeneity, assuming that it varies only with respect to the observed event time. It is not the only possible assumption, and in some applications, other restrictions on  $\alpha(X_i, E_i, e)$  can be appropriate.

We abuse notation and define  $\mu_k(E_i) := \mu_k(E_i, E_i)$ ; we also define  $\nu_{i,k} = \nu_{i,k}(E_i)$ . Straightforward verification implies that

$$\mathbb{E}[\nu_{i,k}|E_i, X_i] = 0. \quad (\text{B.3})$$

Since  $\tilde{W}_{i,k}$  is a deterministic function of  $X_i, E_i$  we can combine all the restrictions and arrive at the model we have in Assumption 4.2:

$$\tau_{i,E_i,k}(\tilde{W}_{i,k}) = \beta_k(X_i) + \mu_k(E_i) + \gamma_{i,k}(\tilde{W}_{i,k}) + \nu_{i,k}, \quad \mathbb{E}[\nu_{i,k}|X_i, E_i, \tilde{W}_{i,k}] = 0.$$