Using Event Studies as an Outcome in Causal Analysis*

Dmitry Arkhangelsky†

Kazuharu Yanagimoto‡

Tom Zohar §

January 27, 2025

Abstract

We propose a causal framework for applications where the outcome of interest is a unit-specific response to events, which first needs to be measured from the data. We suggest a two-step procedure: first, estimate unit-level event studies (ULES) by comparing pre- and post-event outcomes of each unit to a suitable control group; second, use the ULES in causal analysis. We outline the theoretical conditions under which this two-step procedure produces interpretable results, highlighting the underlying statistical challenges. Our method overcomes the limitations of regression-based approaches prevalent in the empirical literature, allowing for a deeper examination of heterogeneity and dynamic effects. We apply this framework to analyze the impact of childcare provision reform on the magnitude of child penalties in the Netherlands, illustrating its ability to reveal nuanced positive relationships between childcare provision and parental labor supply. In contrast, traditional regression-based analysis delivers negative effects, thereby emphasizing the benefits of our two-step approach.

*We thank Manuel Arellano, Stephane Bonhomme, Kirill Borusyak, Clément de Chaisemartin, Nezih Guner, Lance Lochner, and Dean Yang for their feedback, which has greatly improved the paper. We also thank seminar participants at Carlos III, CEMFI, IFAU, and Warwick, and participants in the 15th Research Workshop Banco de España - CEMFI, ELSE 2024, SAE 2024, Econometric Society Winter meeting 2024, Barcelona GSE Summer Forum 2024, and ASSA 2025 for their helpful comments. We gratefully acknowledge financial support from the Maria de Maeztu grant, which funded data access to the Dutch Central Bureau of Statistics and Proyectos de Generación de Conocimiento 2021 for travel expenses to present the project. All views and any errors are our own.

†CEMFI, CEPR, darkhangel@cemfi.es, corresponding author.

‡CEMFI, kazuharu.yanagimoto@cemfi.edu.es.

§CEMFI, CESifo, tom.zohar@cemfi.es.

1 Introduction

Numerous empirical questions focus on how economic outcomes respond to various events, such as changes in the environment (e.g., labor market shocks) or personal decisions (e.g., becoming a parent). To answer these questions, researchers routinely conduct event studies, comparing units that have experienced the event with those that have not yet experienced it. These comparisons are then aggregated to produce a summary that can be easily visualized using event-study plots (e.g., Freyaldenhoven et al., 2021). For many applied questions, this summary is the final object of interest, and no further analysis is necessary. For example, if the original event corresponds to a policy change, the event-study plot describes the policy's impact. However, suppose we are interested in how different policies shape reactions to events. For instance, we want to understand how changes in childcare provision affect changes in individual labor-market outcomes after having children. To answer such questions, we would like to leverage the variation in the underlying unit-level event studies, connecting it to observed policy changes or, in other words, to conduct causal analysis using the event studies as the outcome of interest.

This paper presents a practical framework for systematically exploring unit-specific reactions to events and using them to conduct causal analysis. By connecting results from the econometric panel data literature with advances in the causal inference literature, we provide a toolkit that enables researchers to estimate unit-level reactions and use them to answer causal questions.[1] To demonstrate the practical value of our approach, we apply it to a specific empirical question—the study of child penalties (CPs) and their relationship with childcare provision—and illustrate how it outperforms current methods in this field.

We start from a long-standing observation in the econometric panel data literature (Arellano and Bonhomme, 2012; Bonhomme and Sauder, 2011; Borusyak et al., 2024b; Chamberlain, 1992; Graham and Powell, 2012). Suppose we compare a given unit before and after an event with the average of appropriate control units. As long as the parallel trends assumption holds, these comparisons yield unbiased, though noisy, estimators for unit-level effects, which we refer to as unit-level event studies (ULES). The main proposal of this paper is to directly incorporate these unit-specific measures into the causal analysis, resulting in a two-step procedure: measurement of ULES and the utilization of estimated ULES in the causal analysis.

Many applied researchers already recognize the opportunities presented by the unit-level variation; however, they lack a systematic method for exploring this variation. For instance, a common question in the CP literature focuses on how childcare provision policies impact the magnitude of child penalties (Andresen and Nix, 2022a; Kleven et al., 2024; Lim and Duletzki,

[1] We provide an online tutorial to ease adoption of the two-step approach: https://kazuyanagimoto.com/unitdid/

2023; Rabaté and Rellstab, 2021). Rather than using the two-step procedure we recommend, researchers rely on ad hoc modifications of standard regressions, interacting policy variation with event indicators. This one-step procedure leads to at least three potential issues. First, it limits the types of questions researchers aim to address. For example, as we demonstrate in the context of the CP, conventional regressions do not allow users to differentiate between the contemporaneous and dynamic effects of childcare expansion policies. Second, it constrains the analyses researchers can conduct for a given question and the variation they can examine. Specifically, researchers often arbitrarily discretize the underlying rich policy variation (e.g., low versus high levels of childcare expansion) to be able to use conventional regression specifications. Finally, the current practice can lead to incorrect conclusions. We illustrate this in our empirical application, where the commonly used one-step procedure yields quantitatively and qualitatively different answers compared to our two-step strategy.

In theory, for a given problem, it is often possible to combine the two-step procedure we propose into a single-step regression procedure. However, in practice, there are several advantages to separating the tasks of measuring ULES and conducting causal analysis rather than trying to implement them simultaneously. For example, applied researchers frequently explore various specifications in their analysis. By dividing the analysis into two steps, we can ensure that the empirical strategy in the causal analysis phase does not impact how we measure the ULES. Additionally, some popular techniques depend on statistical methods that involve regularization, which may introduce nonlinearities (e.g., LASSO). Empirical economists frequently use these methods when outcomes are readily available, and our approach enables researchers to apply these same tools in situations with constructed outcomes — ULES. Finally, as we illustrate in the paper, even in the most straightforward problems, designing the appropriate one-step procedure is not trivial, and natural proposals can lead to incorrect answers. Using our approach, researchers can increase the transparency and flexibility of the analysis.

One immediate concern regarding our recommendation is that the constructed unit-specific measures are noisy, and the regression-based analysis that uses such variables may suffer from measurement error bias. We argue that this bias in the measurement step does not impact the causal analysis under standard assumptions. For example, this holds in our empirical application, in which the childcare provision policy varies across municipalities. Intuitively, when analyzing such policies, we need to aggregate unit-level estimates to reflect the level of policy variation, and this aggregation eliminates measurement error in individual-level child penalties.

The second, more nuanced concern is that we can construct ULES only for a selected subsample of units: those that experienced the event. This inherent feature of the analysis leads to sample selection issues, which could undermine causal identification. By dividing the analysis into two steps, we compel researchers to address it explicitly, thereby increasing the trans-

parency and credibility of the analysis. In contrast, conventional regression procedures obscure the sample selection problem, making it easier for researchers to overlook it. We present two alternative assumptions under which the selection concerns can be addressed. The first one assumes away the causal link between the policy of interest and the timing of the event. We argue that this assumption is reasonable within the context of our application and provide empirical evidence to support this claim. However, in applications where this assumption is less reasonable, we offer researchers an alternative restriction, positing that the counterfactual event times and ULES are independent. This assumption is suitable in applications where event times are (quasi) random, such as when determined in an experiment with a policy-dependent design.

We apply our method to study how CPs are affected by a large childcare expansion in the Netherlands. We employ a comprehensive dataset from Statistics Netherlands (CBS) that encompasses various sources, including the census, residential registry, and employment records, all linked through anonymized identifiers by CBS. Our primary data come from employer-employee monthly records from 1999 to 2016, which detail employment statuses and wages. We merge demographic information on birth year, gender, timing of parenthood, and education records for the highest diploma obtained.

We begin our empirical analysis by constructing individual-specific CP estimates. We conduct a basic exploratory analysis to understand the variation in these estimates and demonstrate how to use the estimated ULES to assess the validity of the statistical model. We find supporting empirical evidence for the model's limited anticipation assumption across all combinations of education, gender, and age of first birth, with one exception: college-educated individuals, for whom the model is suitable only for those who have children later in life, beyond the age of 30. Driven by this model validation exercise, we impose this additional sample restriction for the remainder of our analysis.

We continue our analysis by illustrating how ULES can serve as an outcome of interest in causal analysis. Specifically, we examine the effect of the 2005 to 2010 Dutch childcare provision expansion on parent-level estimates of the CP in earnings and employment, using variations in childcare services and availability across municipalities and time. To highlight the significance of the dynamic effects of the policy expansion, we analyze both the level of childcare expansion during the period before childbirth (henceforth "baseline levels") and the levels of contemporaneous childcare provision as the child grows up (henceforth "contemporaneous levels"). While contemporaneous levels of childcare provision increase mothers' earnings, baseline levels impact only higher-educated mothers in the medium term, 3 to 5 years post-birth. In contrast, fathers' earnings and participation are influenced solely by baseline levels. Furthermore, we discover that more educated parents benefit more from the childcare expansion.

These heterogeneous policy effects may provide insights into the economic mechanism underlying the relationship between childcare provision and labor supply. For example, one economic interpretation of these results could be via the role of job flexibility and nonlinear wage structure, which may vary by education level.[2] Childcare expansion disproportionately benefits highly educated women if their jobs have more demanding work-hour requirements (e.g., lawyers), thereby amplifying the impact of childcare availability on their labor market outcomes.

The effect of expansion policies for childcare provision on the CP has recently been studied, with a wide range of results, ranging from null to very large effects (Andresen and Nix, 2022a; Castellanos, 2024; Kleven et al., 2024; Lim and Duletzki, 2023; Rabaté and Rellstab, 2021). Although these varying results might be simply due to different contexts of the reform studied, they may very well be the result of an inflexible empirical specification. Compared with the typical regression analysis in the literature, we approach this question more flexibly by splitting the measurement and policy evaluation steps, which demonstrates the strength of our two-step approach. Typically, both the CPs and their reaction to policies are estimated within a single regression augmented with pre-post policy indicators interacted with binary exposure measures. Such specifications raise several concerns. First, in practice, a policy variation is often continuous (e.g., the increase in childcare provision) rather than binary. Applied researchers often binarize it, thus ignoring essential information we can use directly. Second, common one-step specifications abstract away from dynamic effects, which we find in our applications. Finally, these specifications suffer from aggregation issues and contamination bias (Goldsmith-Pinkham et al., 2024).

Our paper makes both a methodological and empirical contribution. The first step of our analysis—the measurement of ULES—integrates prior results from the econometric panel data literature (Arellano and Bonhomme, 2012; Chamberlain, 1992; Graham and Powell, 2012) with recent studies on event analyses (Borusyak et al., 2024b) to create a simple yet flexible algorithm suitable for all event studies. From an algorithmic perspective, we build on the estimation approach proposed by Borusyak et al. (2024b); however, our measurement framework provides enough flexibility for users to incorporate alternative estimation methods, including those based on Callaway and Sant'Anna (2021) and Sun and Abraham (2021). The second step of our analysis—policy evaluation—combines concepts from modern causal panel data literature (see Arkhangelsky and Imbens (2024) for a recent survey) and the design-based literature (see Borusyak et al. (2024a) for a recent survey) and demonstrates how to apply

[2] Goldin (2014) defines linear jobs as ones that show a proportional relationship between hours worked and wages, allowing for greater flexibility in scheduling. Nonlinear jobs, in contrast, feature convex wage schedules, and impose significant penalties for reducing working hours. The convex structure is commonly a result of a need in specialization, which implies highly educated women are more inclined to pursue them, while lower-educated women tend to remain in linear jobs.

them in situations in which the outcome of interest is not directly available but is measured with error and only for a selected subsample.

The concept of using the two-step procedure to project individual-specific estimates on cross-sectional regressors of interest has a long tradition in panel data econometrics (Arellano and Bonhomme, 2012; Chamberlain, 1992). Our methodological contribution is to expand these ideas to scenarios where the second step aims to uncover a causal relationship, highlighting the challenges inherent in this process. Specifically, we argue that the problems emphasized in the econometric literature—the correlation between unobserved heterogeneity and regressors—play a less significant role in causal analysis, while issues of sample selection and contamination bias become central.

Creating unit-level measures from appropriately transformed data also has a long-standing tradition in causal inference. Early work by Robinson (1988) demonstrated how to use transformed outcomes and regressors to construct efficient estimators in partially linear models. More recently, these concepts have been adapted within the causal inference literature, with applications related to heterogeneous treatment effects and their summaries (Chernozhukov et al., 2018, 2023; Kennedy, 2023; Nie and Wager, 2021; Semenova and Chernozhukov, 2021), as well as policy learning (Athey and Wager, 2021). See Foster and Syrgkanis (2023) for a comprehensive overview and numerous additional references. Our proposal adapts similar ideas to event-study settings and identifies further challenges that must be addressed in this context. Importantly, unlike the existing literature, we use these objects to conduct an additional layer of causal analysis rather than to explore heterogeneity in treatment effects. Compared to prior literature, a key limitation of our statistical analysis is its focus on discrete covariates.

On the empirical side, this paper contributes to the CP literature in several ways. First, it highlights the heterogeneity of CP among individuals, complementing and expanding a body of research that has predominantly focused on aggregate penalties. Previous studies have consistently shown that women experience significant CP after birth, with recovery being partial (Angelov et al., 2016; Gallen, 2019; Kleven et al., 2019). However, emerging literature indicates potential heterogeneity in penalties: Women who express a preference for having a child may not face such penalties (Bensnes et al., 2023; Lundborg et al., 2024), and it has been demonstrated that CP estimation is sensitive to the timing of the first birth, the spacing of births, and parental leave policies (Adams et al., 2024). In response to this concern, we adjust the standard CP estimation to accommodate unobserved heterogeneity and differential trends for subgroups of interest. Second, we explore the mixed evidence regarding the effects of expanding childcare policies on CP (Andresen and Nix, 2022b; Castellanos, 2024; Karademir et al., 2024; Kleven et al., 2024; Lim and Duletzki, 2023; Rabaté and Rellstab, 2021). Our flexible two-step approach helps fill this gap by demonstrating how the common

one-step approach leads to qualitatively different conclusions. Furthermore, we emphasize the importance of dynamics in studying childcare expansion policies.

The remainder of the article is organized as follows. In Section 2, we formally introduce the econometric framework for estimating and using the unit-level measures in causal analysis. Section 3 applies the measurement step to the CP and discusses the data and sample selection. Section 4 illustrates how unit-level estimates can serve as an outcome of interest in policy evaluation exercises, particularly in the context of childcare provision policies, and compares our results from the two-step approach with the common one-step approach found in the literature. Section 5 concludes.

2 Econometric framework

This section outlines the methodological concepts underlying our empirical analysis using a stylized model. The model builds on ideas from previous literature on linear panel data models with random coefficients, particularly Arellano and Bonhomme (2012), as well as more recent studies on heterogeneous treatment effects in event studies, especially Borusyak et al. (2024b). The methodological contribution of this section is to connect these ideas to a broader causal analysis, detailing the conditions under which unit-level estimates can be used for policy evaluation. We introduce the measurement framework, present the causal problem and our solution, and discuss statistical aspects. We opt for a less formal presentation to strike a balance between readability and rigor, with Appendix B providing technical details, formal results, and extended discussions.

2.1 Measurement model

We start with a model that illustrates how to measure the responses of units to a specific event. These measurements, termed "unit-level event studies," will lay the groundwork for the causal analysis outlined in this section.

Consider a population of units observed over three periods, $t \in \{0, 1, 2\}$. For each unit i, we have a set of outcomes $(Y_{i,0}, Y_{i,1}, Y_{i,2})$ and an event time $E_i \in \{1, 2, \infty\}$. Each unit i belongs to a group $g(i)$, which is observable. The event time E_i indicates the period in which the unit experienced a specific event, potentially reflecting a policy change or individual decision depending on the context, with E_i set to ∞ if the event did not occur during the observation period. Depending on the application, group $g(i)$ may represent a geographic location, a classroom, a firm, an economic market, etc. We use a model with only three periods to discuss key challenges in identification, estimation, and inference in the simplest setting; in

Appendix B, we extend the model to allow for multiple periods, event times, and additional covariates.

We assume the outcomes evolve according to a strictly exogenous linear panel data model with unit and group-specific parameters:

$$\begin{aligned} &Y_{i,t} = \alpha_i + \lambda_{g(i),t} + \sum_{h \geq 0} \tau^e_{i,h} \mathbf{1}\{E_i = e\}\mathbf{1}\{t - E_i = h\} + \varepsilon_{i,t}, \\ &\mathbb{E}_{g(i)}[(\varepsilon_{i,0}, \varepsilon_{i,1}, \varepsilon_{i,2}) | \alpha_i, \tau^1_{i,0}, \tau^1_{i,1}, \tau^2_{i,0}, E_i] = 0. \end{aligned} \tag{1}$$

We interpret (1) as a system of measurement equations about the latent parameters of interest, $\tau^1_{i,0}, \tau^1_{i,1}$ and $\tau^2_{i,0}$.[3] Note that $\tau^e_{i,h}$ is indexed by both the event time e, thus allowing for state dependence, and horizon h, thus allowing for dynamics. To simplify notation, we define $\tau_{i,h} := \tau^{E_i}_{i,h}$. The difference between $\tau_{i,h}$ and $\tau^e_{i,h}$ is that the latter object corresponds to the fixed event time e and the former to the random event time E_i. For every unit with $E_i = 1$ and horizon $h \in \{0, 1\}$, we define a unit-level measurement:

$$\hat{\tau}_{i,h} := Y_{i,1+h} - Y_{i,0} - (\lambda_{g(i),1+h} - \lambda_{g(i),0}) = \tau_{i,h} + \nu^1_{i,h},$$

where $\nu^1_{i,h} := \varepsilon_{i,1+h} - \varepsilon_{i,0}$. Similarly for units with $E_i = 2$, we define:

$$\hat{\tau}_{i,0} := \left(Y_{i,2} - \frac{Y_{i,1} + Y_{i,0}}{2}\right) - \left(\lambda_{g(i),2} - \frac{\lambda_{g(i),1} + \lambda_{g(i),0}}{2}\right) = \tau_{i,0} + \nu^2_{i,0},$$

where $\nu^2_{i,0} := \varepsilon_{i,2} - \frac{\varepsilon_{i,0}+\varepsilon_{i,1}}{2}$.[4] The moment restriction in (1) implies that each $\hat{\tau}_{i,h}$ is unbiased for $\tau_{i,h}$. As we will see in the next section, this restriction is key for using $\hat{\tau}_{i,h}$ in causal analysis. For future reference we define the measurement error $\nu_{i,h} := \nu^{E_i}_{i,h}$, which quantifies the error we are making in measuring $\tau_{i,h}$ with $\hat{\tau}_{i,h}$. We refer to $\hat{\tau}_{i,\cdot}$-s, viewed as functions of the horizon h, as unit-level event studies (ULES).

Remark 2.1. We could interpret $\tau^e_{i,h}$ as causal objects by explicitly introducing the underlying potential outcomes and making assumptions that imply the conditional moment restriction; see Arkhangelsky and Imbens (2024) for a discussion of different types of assumptions that guarantee this. Some of these assumptions imply additional properties of the measurement model, which we do not explicitly consider. Instead, we follow Arellano and Bonhomme (2012) and approach the analysis without committing to a specific interpretation of these parameters. Depending on the context, they can be viewed as causal quantities, or (1) can be

[3] The expectation in (1) is taken using the g-specific distribution and conditional on group-level shocks; see Appendix B.1 for the formal definition of the underlying probability model.

[4] In model (1) the differences in the time fixed effects are identified from the differences in outcomes for units that have not experienced the event yet, e.g., $\lambda_{g(i),1} - \lambda_{g(i),0} = \mathbb{E}_{g(i)}[Y_{i,1} - Y_{i,0} | E_i > 2]$.

seen as a statistical model defining the latent parameters.

Remark 2.2. The measurements $\hat{\tau}_{i,h}$ are constructed using specific transformations of the unit-level outcomes $Y_{i,0}, Y_{i,1}, Y_{i,2}$ to deliver an unbiased signal about $\tau_{i,h}$. Other unit-level transformations also achieve the same goal, e.g., we can use a weighted average instead of the simple average of the pre-event outcomes. For instance, for units with $E_i = 2$ we can use $Y_{i,1}$ or $Y_{i,0}$ instead of $\frac{Y_{i,1}+Y_{i,0}}{2}$ to adjust for α_i. This choice affects statistical efficiency, with the simple average being efficient if the variance of $\varepsilon_{i,t}$ is constant over t and the errors are uncorrelated.

2.2 Causal framework

Suppose researchers are interested in a policy intervention W_g that varies at the group level. Such interventions are common in empirical practice, arising in experimental studies in which W_g is randomly assigned at the group level, as well as in observational studies in which implemented policies naturally differ among groups. Unlike the typical policy evaluation problem, the outcome of interest is not directly observed; instead, it must be measured using system (1). For instance, in our empirical example, $\tau^e_{i,h}$ will represent the change in a person's income after they become a parent in period e, while $W_{g(i)}$ will indicate this parent's exposure to a childcare expansion policy.

To formalize the notion of causality, we interpret the observed data as realizations of the underlying potential outcomes (see Imbens and Rubin (2015) for a textbook treatment):

$$Y_{i,t} = Y_{i,t}(W_{g(i)}), \quad E_i = E_i(W_{g(i)}).$$

We connect the potential outcomes $Y_{i,t}(w)$ to the underlying counterfactual latent parameters:

$$Y_{i,t}(w) = \alpha_i(w) + \lambda_{g(i),t}(w) + \sum_{h \geq 0} \tau^e_{i,h}(w) \mathbf{1}\{E_i(w) = e\} \mathbf{1}\{t - E_i(w) = h\} + \varepsilon_{i,t}(w).$$

This notation implies that while the policy intervention affects all system components, it does not change the structure of the measurement model itself.[5] To attach meaning to this decomposition, we impose the moment conditions on the counterfactual quantities:

$$\mathbb{E}[\varepsilon_{i,0}(w), \varepsilon_{i,1}(w), \varepsilon_{i,2}(w) \mid \alpha_i(w), \tau^1_{i,0}(w), \tau^1_{i,1}(w), \tau^2_{i,0}(w), E_i(w)] = 0. \tag{2}$$

Adopting the standard convention, we drop the dependence of w to describe the realized

[5] In some applications, researchers might impose additional restrictions on the causal model. For instance, if W_g describes a path of a policy that evolves over time, then it is natural to assume that $\alpha_i(w)$ does not depend on w.

variables:

$$\alpha_i := \alpha_i(W_{g(i)}), \quad \lambda_{g(i),t} := \lambda_{g(i),t}(W_{g(i)}), \quad \tau^e_{i,h} := \tau^e_{i,h}(W_{g(i)}), \quad \varepsilon_{i,t} := \varepsilon_{i,t}(W_{g(i)}).$$

We use the fact that $W_{g(i)}$ varies at the group level and the moment restriction (2) to arrive at model (1), which allows us to define $\hat{\tau}_{i,h}$ in the same way as before.

We now introduce the main counterfactual object of our causal analysis:

$$\tau_{i,h}(w) := \tau^{E_i(w)}_{i,h}(w).$$

In principle, this quantity is affected by w along two distinct margins. First, there is a direct effect of w on $\tau^e_{i,h}(w)$ for a given event time e and horizon h. Second, there is an indirect effect of w through the changes in the potential event time $E_i(w)$. For example, expanding the childcare provision might affect both the child penalty and the timing of having a child. Following the notation introduced in the previous section, we use $\tau_{i,h}$ to denote the realized potential outcome:

$$\tau_{i,h} := \tau_{i,h}(W_{g(i)}).$$

The properties of $\hat{\tau}_{i,h}$ ensure that it provides an unbiased signal for $\tau_{i,h}$, with the measurement error $\nu_{i,h}$ being uncorrelated with any function of $W_{g(i)}$ by construction.[6] This suggests using ULES for causal analysis, which we focus on in the next section.

2.3 Identification

Once we have constructed the ULES, we can use these quantities as the outcomes of interest in causal analysis, which is the primary focus of this article. The specific nature of this exercise should depend on the assumptions regarding the variation in W_g. For instance, if researchers control the assignment of W_g, the analysis will be significantly simplified. However, we must confront the unavoidable sample selection problem even in such cases. By definition, we observe $\hat{\tau}_{i,h}$ only for selected subpopulations; specifically, we can construct $\hat{\tau}_{i,0}$ only for units with $E_i < \infty$. This issue becomes even more critical when comparing units with different event times—a necessity in observational studies where researchers do not control the assignment of W_g.

To resolve the sample selection problem, we restrict the underlying causal model. We consider an assumption that allows for two options:

$$\text{either (a)} \quad E_i(w) \equiv E_i, \quad \text{or (b)} \quad E_i(w) \perp\!\!\!\perp \tau^e_{i,h}(w) \text{ within group } g(i). \tag{3}$$

[6] This is guaranteed because the errors $\varepsilon_{i,t}(w)$ have zero mean within each group; see Appendix B.1 for technical details.

Restriction (a) is natural in applications in which event times are not expected to respond to the policy of interest. This assumption is plausible if the observed distribution of E_i remains unchanged in relation to $W_g(i)$, which—as we will argue in Section 4.3—is the case in our empirical application. This restriction automatically resolves the selection problem because E_i is not affected by the policy.

A large class of applications involves event times responding to policies, which leads us to consider an alternative restriction (b) that eliminates the dependence between $E_i(w)$ and $\tau^e_{i,h}(w)$ at the group level. In this scenario, while the selection problem persists, it can be addressed by conducting appropriate empirical analysis. This assumption is automatically fulfilled if $E_i(w)$ is assigned randomly according to a group-specific distribution, as in a randomized experiment.

This discussion highlights a crucial practical benefit of separating measurement from causal analysis. The explicit construction of ULES brings the selection problem to the forefront, forcing researchers to confront it. Conventional one-step methods that bypass the explicit measurement step obscure the selection issue, diminishing transparency and potentially leading to incorrect conclusions.

Remark 2.3. Assumption (3) covers a broad range of applications; however, there are empirical problems for which it is unreasonable. In such situations, researchers have multiple options. One option is to conduct an unconditional analysis—for example, by altering the outcome of interest from $\tau_{i,0}$ to $\tau_{i,0}\mathbf{1}\{E_i < \infty\}$. An alternative approach would be to address the selection problem directly by either constructing bounds, as outlined by Lee (2009) or by imposing additional structure on the joint distribution of $E_i(w)$ and $\tau_{i,h}(w)$ using a correlated random effects model. The latter methods necessitate further assumptions, which we do not elaborate on in this paper.

Known distribution of W_g

To see the implications of Assumption (3) more formally, suppose that $W_g \in \{0, 1\}$ and is assigned randomly across groups with probability $\frac{1}{2}$, which is known to the researcher. In this case, we can rely on inverse probability weighting at group level to conduct the causal analysis. Specifically, we need to multiply group-level outcomes by the weights $\frac{W_g}{0.5} - \frac{1-W_g}{0.5} = (4W_g - 2)$. This leads to the following two computations. If part (a) of (3) is satisfied, then we have

$$\mathbb{E}\left[\frac{(4W_{g(i)} - 2)\hat{\tau}_{i,0}\mathbf{1}\{E_i < \infty\}}{\mathbb{E}_{g(i)}[\mathbf{1}\{E_i < \infty\}]}\right] = \mathbb{E}[(4W_{g(i)} - 2)\tau_{i,0}(W_{g(i)})|E_i < \infty] = \mathbb{E}[\tau^1_{i,0}(1) - \tau^1_{i,0}(0)|E_i < \infty].$$

Here the first equality follows from the fact that when $\hat{\tau}_{i,0}$ is available—hence conditioning on $E_i < \infty$—it is an unbiased estimator of $\tau_{i,0}(W_{g(i)})$ for each group. The second equality follows from the random assignment of $W_{g(i)}$ across groups.

Alternatively, if part (b) of (3) is satisfied, then we have for $e \in \{1, 2\}$

$$\mathbb{E}\left[\frac{(4W_{g(i)} - 2)\hat{\tau}_{i,0}\mathbf{1}\{E_i = e\}}{\mathbb{E}_{g(i)}[\mathbf{1}\{E_i = e\}]}\right] = \mathbb{E}\left[\frac{(4W_{g(i)} - 2)\tau^e_{i,0}(W_{g(i)})\mathbf{1}\{E_i(W_{g(i)}) = e\}}{\mathbb{E}_{g(i)}[\mathbf{1}\{E_i(W_{g(i)}) = e\}]}\right] = \mathbb{E}[(4W_{g(i)} - 2)\tau^e_{i_0}(W_{g(i)})] = \mathbb{E}[\tau^e_{i,0}(1) - \tau^e_{i,0}(0)].$$

In this case, the first equality again uses the unbiasedness property of $\hat{\tau}_{i,0}$, the second follows from $E_i(w)$ being independent of other potential outcomes within the group, and the last one relies on W_g being randomly assigned.

These computations illustrate two important differences between the restrictions in (3). In the first case, the analyst does not need to fix a particular event time to compute the policy's impact.[7] However, the resulting estimand describes the effects for a selected subpopulation within each group, similar to the average effect on the treated. In the second case, the analysis must be done separately for each event time to account for selection. Meanwhile, the derived effect corresponds to the average treatment effect for the entire group. As we will show below, the two restrictions also have different implications for statistical estimation.

The above analysis can be generalized to environments in which W_g is randomly assigned based on observed group-level covariates or has a more complicated non-binary structure. As long as Assumption (3) is satisfied and the distribution of W_g can be learned, researchers can use the machinery developed by modern cross-sectional causal inference literature, including recent proposals such as the automatic debiased machine learning described by Chernozhukov et al. (2022). Similarly, we can use IV or regression-discontinuity identification strategies, which rely on between-group comparisons.

Unknown distribution of W_g

In empirical practice, it is common for policy variation not to be controlled by the researcher and to have an unknown group-specific distribution. In such situations, we cannot disentangle the effect of the policy from the spurious correlation in the data without additional assumptions. This issue can be resolved only on a case-by-case basis, and the approach we present below is motivated by our application. Nevertheless, we believe it is flexible enough to address many empirical problems. Importantly, it also generalizes the one-step regressions used in practice, which we discuss in Section 4.4.

[7] The first computation remains valid for a fixed value of $E_i = e$, which results in a different estimand $\mathbb{E}[\tau^1_{i,0}(1) - \tau^1_{i,0}(0)|E_i = e]$.

Suppose $W_{g(i)} = (W_{g(i),0}, W_{g(i),1}, W_{g(i),2})$—that is, each group has a particular policy path. In our application, each $W_{g,t}$ describes the level of childcare provision in municipality g in calendar period t. We assume that only part of the policy path is relevant to the current potential outcomes and the effect is linear:

$$\tau^e_{i,h}(w_0, w_1, w_2) = \beta^e_{i,0} + \delta_{base,0} w_{e-1} + \delta_{cont,0} w_{e+h}. \tag{4}$$

Here, w_{e-1} and w_{e+h} correspond to the baseline level of the policy before the event and the current level, respectively. Restriction (4) implies that the effect of past policies is homogeneous across time and groups. This assumption can be relaxed at the expense of additional restrictions, which we discuss at length in Appendix B.3. Note that we leave β^e_i completely unrestricted, and it can systematically vary over groups and event times. Below we focus on $\tau_{i,0}$ because we can construct $\hat{\tau}_{i,0}$ for units with various event times, unlike $\hat{\tau}_{i,1}$. In our empirical application, we consider all horizons.

Next, we restrict the probability model for $W_{g,t}$ and assume

$$\mathbb{E}[W_{g,t}] = a_g + b_t. \tag{5}$$

This assumption is reminiscent of those invoked in the recent design-based IV literature (e.g., Borusyak and Hull, 2023, 2024; Borusyak et al., 2024a), as well as the recent literature on linear regression (Goldsmith-Pinkham et al., 2024). It allows us to use the variation in differences $\Delta W_{g,t} := W_{g,t} - W_{g,t-1}$ as an instrument because this variation does not have a systematic group-specific component. The group-specific component of $\mathbb{E}[W_{g,t}]$, a_g, is eliminated by the transformation $\Delta \boldsymbol{W}_{g(i)}$, and the common mean $\mathbb{E}[\Delta \boldsymbol{W}_{g(i)}] = b_2 - b_1$ can be consistently estimated by pooling the data across groups.

To understand the underlying mechanics, define

$$\Delta\hat{\tau}_{i,0} := \frac{\hat{\tau}_{i,0}\mathbf{1}\{E_i = 2\}}{\mathbb{E}_{g(i)}[\mathbf{1}\{E_i = 2\}]} - \frac{\hat{\tau}_{i,0}\mathbf{1}\{E_i = 1\}}{\mathbb{E}_{g(i)}[\mathbf{1}\{E_i = 1\}]},$$

and $\Delta \boldsymbol{W}^\top_{g(i)} := (W_{g,2} - W_{g,1}, W_{g,1} - W_{g,0})$. Observe that we have

$$\begin{aligned}
&\mathbb{E}\left[\Delta\tau_{i,0}\left(\Delta \boldsymbol{W}_{g(i)} - \mathbb{E}[\Delta \boldsymbol{W}_{g(i)}]\right)\right] = \\
&\mathbb{E}\Bigg[\left(\Delta \boldsymbol{W}_{g(i)} - \mathbb{E}[\Delta \boldsymbol{W}_{g(i)}]\right)\left(\left(\frac{\beta^2_{i,0}\mathbf{1}\{E_i = 2\}}{\mathbb{E}_{g(i)}[\mathbf{1}\{E_i = 2\}]} - \frac{\beta^1_{i,0}\mathbf{1}\{E_i = 1\}}{\mathbb{E}_{g(i)}[\mathbf{1}\{E_i = 1\}]}\right) + \right. \\
&\left.\Delta \boldsymbol{W}^\top_{g(i)}\boldsymbol{\delta}_0\right)\Bigg] = \mathbb{V}[\Delta \boldsymbol{W}_{g(i)}]\boldsymbol{\delta}_0,
\end{aligned}$$

where $\boldsymbol{\delta}_0^\top := (\delta_{cont,0}, \delta_{base,0})$. Combining the above moment across groups, we can construct a consistent estimator for $\boldsymbol{\delta}_0$. The next section will discuss a particular regression estimator that implements this strategy.

Remark 2.4. Restriction (5) is quite specific and may not be suitable for all applications. In Appendix B.3, we discuss alternative restrictions on the mean, particularly those that allow for a factor model. The overall logic remains the same: As long as there exists an identifiable transformation of the policy path such that its mean is invariant across groups, we can use it for identification.

Remark 2.5. The restriction (5) is reminiscent of assumptions on the propensity score that are common in the literature on design-based causal identification. At the same time, in the policy evaluation literature, it is common to use the model-based approach, keeping the moments of $W_{g,t}$ unrestricted while imposing structure on the residual term $\beta^e_{i,h}$. We discuss this alternative strategy in Appendix B.5. We argue that our estimator remains valid under a specific version of the parallel trends assumption (analogous to the one used by De Chaisemartin and d'Haultfoeuille, 2020), even if restriction (5) fails. This double robustness identification property is closely related to the results established by Arkhangelsky and Imbens (2022); Arkhangelsky et al. (2024a); see also Borusyak and Hull (2024).

2.4 Statistical analysis

This section explores the practical implementation of the ideas discussed earlier. We cover estimators for the measurement step, policy analysis, and the validation of the underlying models.

Measurement

The measurement model in Section 2.1 is overidentified, and thus we can use various procedures for estimation. In our empirical application that features multiple treatment periods, we rely on the imputation estimator developed by Borusyak et al. (2024b), which we review in Appendix B.4. In the context of the example in this section, this estimator has a straightforward structure for $\hat{\tau}_{i,0}$ that reduces to the standard difference-in-difference (DiD) estimator at group level:

$$\hat{\tau}^{BJS}_{i,0} = \left(Y_{i,E_i} - \frac{1}{E_i}\sum_{l<E_i} Y_{i,l}\right) - \frac{\sum_{j:E_j>E_i, g(j)=g(i)}\left(Y_{j,E_i} - \frac{1}{E_i}\sum_{l<E_i} Y_{j,l}\right)}{\sum_{j=1}^n \mathbf{1}\{g(j)=g(i), E_j > E_i\}},$$

which can be constructed for units with $E_i < \infty$.[8] We can immediately see the connection between $\hat{\tau}^{BJS}_{i,0}$ and $\hat{\tau}_{i,0}$:

$$\hat{\tau}^{BJS}_{i,0} = \hat{\tau}_{i,0} + \frac{\sum_{j:E_j>E_i, g(j)=g(i)} \left(\varepsilon_{j,E_i} - \frac{1}{E_i} \sum_{l<E_i} \varepsilon_{j,l} \right)}{\sum_{j=1}^{n} \mathbf{1}\{g(j) = g(i), E_j > E_i\}} = \hat{\tau}_{i,0} + \xi_{i,0}, \quad \mathbb{E}_{g(i)}\left[\xi_{i,0} | E_i\right] = 0.$$

We will refer to $\xi_{i,h}$ as the estimation error to differentiate it from the measurement error $\nu_{i,h}$ discussed earlier. Unlike the measurement error, the estimation error is negligible when the number of relevant units in the group is large. Since we focus on the analysis where $\hat{\tau}^{BJS}_{i,h}$ serves as the dependent variable, these errors do not introduce bias. However, this changes if we intend to use $\hat{\tau}^{BJS}_{i,h}$ as an independent variable, where the measurement error becomes important for identification.[9]

The computation above highlights an additional issue that we ignored in the identification analysis of the previous sections. For a given group g, we can estimate the ULES only if there is variation in event times within the group—an overlap condition. If all the groups are large, such variation will likely exist for each of them. However, if the groups contain only a few units, then the overlap is likely to fail for a significant portion of the groups. Whether the failure of overlap poses a problem depends on which aspect of Assumption (3) we depend on. Suppose the analysis relies on part (a); in that case, we can disregard the groups for which we cannot construct $\hat{\tau}^{BJS}_{i,0}$, and this type of sample selection does not introduce additional challenges for the policy evaluation step, although it does influence the estimand we can construct. However, if the analysis depends on part (b) of Assumption (3), then focusing on the groups for which the overlap holds introduces a group-level sample selection problem acting as a "bad control" (Angrist and Pischke, 2009).

We want to emphasize that this issue is relevant only when the groups are small. For a given group, the probability of overlap failing decreases exponentially with the number of units within the group. This guarantees that the selection problem is statistically inconsequential if the number of units in each group is sufficiently large.

Policy analysis

Once estimated ULES are available, researchers can use them for policy analysis. The nature of this exercise depends on the variation we can employ and the type of assumptions we are willing to accept regarding the underlying causal model. In our discussion below, we focus on the case of an unknown distribution of W_g, while assuming the restrictions introduced in the

[8] See Arkhangelsky and Samkov (2024) for the relevant derivations.

[9] The earlier version of this paper, Arkhangelsky et al. (2024b), which is available on arXiv, discusses this in detail.

previous section. Our general results (see Appendix B.3.1) can be used to construct estimators for applications where the distribution of W_g is known.

Suppose we assume that part (a) of Assumption (3) holds. Then researchers can use the following estimator:

$$(\{\hat{\mu}_g^{OLS}\}_g, \{\hat{\beta}_e^{OLS}\}_e, \hat{\delta}_{base}^{OLS}, \hat{\delta}_{cont}^{OLS}) := \\ \arg \min_{\{\mu_g\}_g, \{\beta_e\}_e, \delta_{base}, \delta_{cont}} \sum_i \left(\hat{\tau}_{i,0}^{BJS} - \mu_{g(i)} - \beta_{E_i} - \delta_{base} W_{g(i),E_i-1} - \delta_{cont} W_{g(i),E_i}\right)^2, \quad (6)$$

where μ_g and β_e are the group and event-time fixed effects, respectively. These two-way fixed effects automatically adjust for the group-level variation in the mean of $W_{g,t}$, making $(\hat{\delta}_{base}^{OLS}, \hat{\delta}_{cont}^{OLS})$ a consistent estimator of the coefficient of interest $\boldsymbol{\delta}_0$.

Alternatively, if part (b) of (3) holds, then researchers need to use a weighted regression:

$$(\{\hat{\mu}_g^{WOLS}\}_g, \{\hat{\beta}_e^{WOLS}\}_e, \hat{\delta}_{base}^{WOLS}, \hat{\delta}_{cont}^{WOLS}) := \\ \arg \min_{\mu_g, \beta_e, \delta_{base}, \delta_{cont}} \sum_i \left(\hat{\tau}_{i,0}^{BJS} - \mu_{g(i)} - \beta_{E_i} - \delta_{base} W_{g(i),E_i-1} - \delta_{cont} W_{g(i),E_i}\right)^2 \frac{1}{\hat{\pi}_{g(i)}(E_i)},$$

where $\hat{\pi}_{g(i)}(E_i) := \frac{\sum_{j:g(i)=g(j)} \mathbf{1}\{E_g = E_i\}}{\sum_{j=1}^n \mathbf{1}\{g(j)=g(i)\}}$. The weighting is needed to account for the sample selection problem caused by the effect of $W_{g(i)}$ on $\hat{\pi}_{g(i)}(E_i)$.

In our empirical analysis, we rely on a generalization of (6) that accommodates unit-specific covariates and multiple horizons h, which we detail and analyze formally in Appendix B.4. We make this choice after conducting preliminary analyses suggesting that part (a) of Assumption (3) is reasonable for our dataset. We report conventional standard errors, clustering at the group level. In Appendix B.4, we formally demonstrate that this approach ensures correct asymptotic inference despite measurement and estimation errors in the left-hand-side variable. This guarantees that researchers can treat the estimated $\hat{\tau}_{i,h}^{BJS}$ as a typical outcome variable in policy analysis while ignoring the estimation error.

Validation

We rely on the overidentifying restrictions in (1) to validate the measurement model. In particular, we have the conventional parallel trends restriction:

$$\mathbb{E}_{g(i)}[Y_{i,1} - Y_{i,0} | E_i = 2] = \mathbb{E}_{g(i)}[Y_{i,1} - Y_{i,0} | E_i = \infty].$$

We can use this population restriction to define a group-specific error:

$$\hat{\nu}_g^1 := \frac{\sum_{i:g(i)=g,E_i=2}(Y_{i,1}-Y_{i,0})}{\sum_{i=1}^n \mathbf{1}\{g(i)=g,E_i=2\}} - \frac{\sum_{i:g(i)=g,E_i=\infty}(Y_{i,1}-Y_{i,0})}{\sum_{i=1}^n \mathbf{1}\{g(i)=g,E_i=\infty\}}.$$

Under the model (1), this error has a zero mean as long as it is well defined, i.e.,

$$\mathbb{E}_g\left[\hat{\nu}_g^1 | \sum_{i=1}^n \mathbf{1}\{g(i)=g,E_i=\infty\}>0, \sum_{i=1}^n \mathbf{1}\{g(i)=g,E_i=2\}>0\right]=0.$$

This condition provides empirical researchers with one testing restriction per group to validate the measurement model. Regardless of which part of Assumption (3) holds, it remains valid. In our empirical application, we validate the model by looking at averages of $\hat{\nu}_g^1$ defined by observed covariates; see Section 3.2 for details.

Remark 2.6. As long as the number of units in each group is large, $\hat{\nu}_g^1$ is approximately normal, and we can use the conventional t-statistic at the group level (with the standard error estimated by within-group bootstrap or analytically). When, in addition, the number of groups is large, all $\hat{\nu}_g^1$ will be approximately jointly normal as long as the conditions of the appropriate high-dimensional CLT hold. In this case, we can use aggregate statistics, such as $\max_g |\hat{\nu}_g^1|$, to test the moment restriction. We can rely on multiplier bootstrap to compute the corresponding critical values. See Chernozhukov et al. (2017) for details and technical requirements for the high-dimensional CLT.

2.5 Discussion

In this section, we discuss three key aspects of our approach. First, we emphasize its advantages over one-step procedures that do not differentiate between measurement and causal analysis. Second, we examine the significance of group-level policy variation, contrasting it with instances where this variation occurs at the unit level. Finally, we discuss our results within the context of the econometric literature, underscoring the additional nuances introduced by our emphasis on causal analysis.

One- vs. two-step approach

The construction of $\hat{\tau}_{i,h}^{BJS}$ is based on linear combinations of the observed outcomes. Moreover, most methods for causal analysis are also linear in the outcome variable. Since the combination of two linear procedures is linear, the two-step analysis can be implemented in a single step, using weighted OLS or two-stage least squares, depending on the exact nature of the empirical exercise. Practitioners often prefer a more straightforward one-step procedure, but, as we

argue below, separating them can be more prudent, transparent, and flexible in many empirical applications.

There are two reasons for the added flexibility of the two-step approach. First, in a typical empirical application, researchers use different identification strategies and regression specifications for causal analysis to guarantee the robustness of the results. In this case, each robustness check would require a different one-step procedure, adjusting the specification in ways that do not always follow standard practice (see our example in Section 4.4). Under the two-step procedure we advocate, the measurement step is unrelated to the identification argument behind the causal analysis, which renders the overall analysis more transparent. Moreover, separating the process into two steps does not affect the construction of standard error; researchers can continue using the same clustering methods they would have used otherwise. Thus, there is essentially no practical cost in using our two-step approach. Second, the methodological literature constantly produces new identification frameworks and estimators for causal analysis problems. By using our two-step approach, these frameworks and estimators can be deployed directly without any complications.

Finally, using the two-step procedure is more prudent since constructing the correct one-step procedure can be less straightforward than one might think. To illustrate the last problem in the simplest setting, suppose that researchers are interested in the effect of a binary policy variable $W_g \in \{0, 1\}$, which was randomly assigned at the group level with probability $\frac{1}{2}$. A natural one-step procedure that comes to mind for such a setting would be to estimate the following linear equation by OLS with fixed effects:

$$Y_{i,t} = \alpha_i + \lambda_{g,t} + \sum_{h \geq 0} (\beta_h + \delta_h W_{g(i)}) \mathbf{1}\{t - E_i = h\} + \epsilon_{i,t}, \tag{7}$$

interpreting the resulting OLS estimator $\hat{\delta}_h$ as the causal effect of the policy. Given that W_g is randomly assigned, one might expect this estimator to deliver a meaningful causal effect.

In Appendix B.7, we demonstrate that if the measurement equation (1) is correctly specified and Assumption (3) holds, the estimator $\hat{\delta}_h$ (7) may not have a meaningful causal interpretation even when the policy is randomly assigned.[10] The reasons for this failure depend on which part of Assumption (3) we rely on. In the case of part (a), the resulting estimator is affected by contamination bias as introduced by Goldsmith-Pinkham et al. (2024). Consequently, as long as the underlying policy effects are heterogeneous, the resulting estimator may lack a meaningful causal interpretation. Importantly, this is the case as long as any type of heterogeneity—across groups, event times, or calendar time—is present.

[10] To simplify the exposition, we consider an estimator that uses the difference $Y_{i,t} - Y_{i,0}$ to eliminate the unit-level fixed effects rather than the event-time specific transformation. We anticipate similar results for the standard OLS estimator.

In the case of part (b) of Assumption (3), the problem is more severe, and $\hat{\delta}_h$ is invalid even when there is no heterogeneity in the underlying effects of the policy. The issue arises because when estimating (7) we use the event-time indicators, which serve as "bad controls" introducing sample selection bias. This discussion shows that natural one-step solutions can backfire even in the most straightforward settings. In more complicated observational studies, researchers will likely consider a more elaborate version of regression (7), potentially introducing additional biases.

Group- vs. individual-level policy variation

Our analysis centered on policies that differ at the group level, a common scenario in many empirical applications. These groups can vary significantly in nature, representing geographic locations, firms, markets, and so on, with sizes ranging from thousands to just a few units. Our framework accommodates all these situations. However, in some cases, policy variation genuinely happens at the unit level. For instance, in the context of child penalties, one might consider directly allocating childcare subsidies to families, creating variation at the individual level that does not align with any specific group. Therefore, understanding the distinctions between the two settings is crucial.

Suppose the policy W_i is randomly assigned at the individual level, and we observe many randomly sampled units. An appropriate version of the measurement model (1) for this situation has the following form:

$$Y_{i,t} = \alpha_i + \lambda_t(W_i) + \sum_{h \geq 0} \tau^e_{i,h} + \varepsilon_{i,t}, \quad \mathbb{E}[\varepsilon_{i,t} | W_i, E_i, \alpha_i, \tau^1_{i,0}, \tau^0_{i,1}, \tau^2_{i,0}] = 0.$$

This model explicitly relies on the policy W_i in question. Consequently, one must either consider separate measurement systems for different policies or start with a comprehensive set of policies. The first approach presents conceptual challenges, while the second is impractical. Naturally, one could bypass this issue by assuming that λ_t does not vary with W_i, but this significantly limits the underlying causal model. This problem does not arise with policies that vary across groups, as we can always incorporate group-level fixed effects, thereby explicitly accounting for any group-level policy.

The previous discussion highlights the complexities of unit-level variation, but one might wonder if it offers any advantages, in particular, if the natural one-step methods now perform better. In Appendix B.7 we consider estimation of a linear equation

$$Y_{i,t} = \alpha_i + \lambda_t(W_i) + \sum_{h \geq 0} (\beta_h + \delta_h W_i) \mathbf{1}\{t - E_i = h\} + \epsilon_{i,t}, \tag{8}$$

using OLS. We demonstrate that the resulting estimator $\hat{\delta}_h$ generally lacks causal interpretation, suffering from selection issues and contamination biases.

The final distinction between the two cases lies in the approach to inference. With group-level policy variation, one can disregard measurement and estimation errors and rely on group-level clustering. This presents a significant practical advantage, allowing users to treat estimated quantities as data. However, this logic is no longer valid when the policy varies across units. In such instances, we must explicitly address the correlation in $\hat{\tau}_{i,h}^{BJS}$ introduced by the estimation errors. Depending on the application, this can be straightforward, such as when the analysis can be completed in one step, as previously described, or it may require additional computations.

Given our focus on group-level variation, one could argue that there is no need to estimate unit-level event studies (ULES); instead, we can directly measure group-level parameters. We focused on the unit-specific analysis for three reasons. First, unit-level analysis is important when units differ in their observed characteristics, and we need to control for that in the causal analysis. This flexibility comes at no practical cost, making it an attractive default option. Second, by using unit-level outcomes in regressions, we automatically weigh group-level quantities according to the number of units in each group—, a commonly used strategy with group-level data. Finally, ULES allows us to study persistence patterns in $\tau_{i,h}$ at the expense of additional assumptions.[11] While we do not attempt such analysis in our empirical application, other researchers might find it useful.

Causal analysis vs. projections

Existing econometric literature shows that conventional one-step regressions do not recover meaningful parameters in models with unobserved heterogeneity, such as our measurement model (1). This argument was presented by Chamberlain (1992), who focused on estimating the mean, and was further elaborated by Arellano and Bonhomme (2012), who investigated the estimation of arbitrary projections of unobserved heterogeneity. The econometric concerns regarding one-step regressions are well-understood: The unobserved heterogeneity may be correlated with the regressors; for example, $\tau_{i,h}^{e}$ could be correlated with E_i. By conducting the analysis in two steps, researchers explicitly address this correlation. In contrast, the one-step regressions push the unobserved heterogeneity into the error term, leading to an endogeneity issue; see Muris and Wacker (2022) for a formal comparison.

Our causal analysis provides a different perspective on this issue, enabling us to interpret the resulting estimands. As discussed in the previous two sections, the one-step approach typically does not yield a meaningful causal quantity. However, the reasons for this are more

[11] See the previous version of this paper for particular assumptions.

nuanced than highlighted in traditional econometric literature. In particular, selection and contamination biases play a significant role. In simpler models, these issues may not occur, and the one-step procedure can produce an interpretable causal quantity. We illustrate this using a simple example in Appendix B.7.

This discussion highlights two key points. First, it is essential to specify the underlying causal model when conducting policy analysis on the measured unit-level parameters. To our knowledge, this paper is the first to explicitly tackle this issue and underscore the associated challenges. Second, the well-known econometric problems with one-step approaches may be less pertinent in the context of causal analysis. Instead, it is important to confront selection problems, address contamination bias, and clarify which type of policy evaluation identification strategy to employ. By separating the measurement and evaluation steps, researchers can effectively address all these issues.

3 Measuring individual-level child penalties

In this section, we measure the ULES described in Section 2.1 in a particular empirical problem: the estimation of child penalties (CPs).[12] We begin by describing the data we use for this purpose. We then describe the construction of ULES—which, in this particular context, we also call individual-level CP—and explore the observed and unobserved heterogeneity in these quantities. These data and the individual-level CP measures are then used in Section 4, in which we analyze the effects of childcare expansion policies.

3.1 Data

Data Sources

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

Tax and Employment Records Our primary data source is an extensive annual-level employer-employee data set derived from tax records (*baansommentab*) covering 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,

[12] In their influential paper, Kleven et al. (2019) coined the term "child penalty" to describe the differential career and earnings losses incurred by women compared with men after having children.

excluding self-employment. Second, earnings data consist 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 birth year, date of death, sex, and annual 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 birthdates and the linkage between parents and their children to determine the first child for all legal parents, which may include both adoptive and 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, which we classify into three levels: high school, vocational training, and bachelor's degree. We exclude individuals with higher education (MA and PhD) and lower education (below high-school subpopulations for two reasons: (a) they form a smaller share of the population, and (b) fertility and labor market decisions will likely follow a different pattern in those groups.[13]

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 on 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 (given from *gemstplaatsbus/gemtplbus/ngemstplbus*). The job location data set contains each worker's municipality and firm ID, which we merge with the firm classification data.

Sample Definition

A key aspect of our study is the examination of labor market outcomes around the time of first childbirth. We restrict the sample to individuals born in 1993 or earlier to ensure we observe labor market outcomes at sufficiently mature ages. To capture transitions into parenthood, we include only those whose age at first birth was below 44, as observed from 2003 onward. To ensure adequate labor market attachment before parenthood, we further restrict the sample to individuals who became parents at least 6 years after the typical graduation age for their highest educational attainment: 24 years for high school graduates, 26 years for vocational degree holders, and 27 years for those with a bachelor's degree. This approach provides a balanced panel of pre-parenthood labor market trajectories while minimizing censoring concerns.

[13] These different life-cycle earnings patterns translate to violations of our statistical model, requiring another sample restriction.

3.2 Measurement

To estimate individual CP, we follow the approach outlined in Section 2.1 describing the general estimation of ULES. For each individual, we observe several covariates: gender $G_i \in \{\text{male; female}\}$, education level $educ_i$, and birth cohort B_i; we collect these covariates into a vector $X_i = (G_i, educ_i, B_i)$ and conduct the analysis separately for each possible value of X_i. The groups $g(i)$ in our application correspond to municipalities, which we discuss in more detail in the next section.

Compared with the simple model outlined in Section 2.1, we need to make several changes to incorporate the additional complexity of the empirical application. The first adjustment is conceptually straightforward: We observe individuals for many years and thus have many different event times. The conceptual extension of the measurement model to allow for multiple periods before and after the event is immediate, and we explain the practical implementation in Appendix B.4.

The second adjustment is motivated by the economic forces behind the observed data. The measurement model (1) is based on the concept of an event time, which is not clearly defined in the context of childbirth. For instance, individuals can change their economic decisions today in anticipation of having children in the future. Moreover, the observed childbirth is a realization of the underlying latent process of individuals attempting to have children. Therefore, to make the measurement system (1) meaningful, we need to consider these anticipation and uncertainty concerns. We opt for a more agnostic and data-driven approach rather than committing to a particular dynamic model behind the observed data. Specifically, we allow childbirth to affect labor market outcomes for up to 3 years before the event. We then test and find empirical support for the resulting model using the methods outlined in Section 2.4. From an algorithmic standpoint, this amounts to shifting the observed childbirth year by 3 years and then applying the approach described in Section 2.1 and Appendix B.4.

We estimate $\hat{\tau}_{i,h}$ with $h \in \{-3, \dots, 5\}$, where negative values of h correspond to pre-event periods. Our approach is similar to the conventional one used in the literature on CPs (e.g., Kleven et al., 2019), but with two important distinctions. First, we use a larger control group by considering the outcomes of *all* individuals who have children later in life, not just those who have children 1 year apart. Second, we adjust for permanent differences in pre-event outcomes to account for possible compositional differences among groups of individuals who have children at different ages. We apply the imputation algorithm of Borusyak et al. (2024b) to construct estimates $\hat{\tau}^{BJS}_{i,h}$ for $\tau_{i,h}$.[14]

In line with Remark 2.1, we do not commit to an interpretation of $\tau_{i,h}$ as "causal effects"

[14] As discussed in Remark 2.2 and Section 2.4, there are multiple valid estimators for $\tau_{i,h}$. In particular, our methodology allows users to choose a more restricted control group, such as individuals who had children in the subsequent year. We use the imputation estimator because we expect it to be more efficient.

of childbirth, but rather view the measurement system we use to construct estimates for $\tau_{i,h}$ as effectively defining CPs as latent variables of interest. Attaching causal meaning to $\tau_{i,h}$ is challenging for several reasons. As discussed above, childbirth, or lack of it, results from many unobserved decisions individuals make. One approach to this problem is relying on explicit randomness that affects these decisions (e.g., birth control failure or reproductive medicine success).[15] When feasible, this strategy recovers LATE-type parameters relevant only for particular subpopulations.

In contrast, by leveraging the measurement model (1), we can calculate $\hat{\tau}_{i,h}^{BJS}$ for a significant share of individuals with children. This is only useful if the comparisons we use to estimate $\hat{\tau}_{i,h}^{BJS}$ are reasonable or, formally, if the measurement model (1) is correct. Therefore, we use the testable implications discussed in Section 2.4 to validate them. Specifically, we can observe how well we adjust for differences across various groups by examining $\hat{\tau}_{i,h}^{BJS}$ for negative values of h. Indeed, we identify violations of this validation exercise for some combinations of education groups and age at first birth, which motivates our further sample selection. We elaborate on the results of these diagnostic tests in more detail below.

Since CPs are more naturally interpreted in relation to underlying baseline outcomes (earnings or participation), we normalize them using average predicted outcomes based on model (1); see Appendix B.4.4 for details. For future reference, we introduce the following notation for the normalized ULES:

$$\tilde{\boldsymbol{\tau}}_i^{BJS} := \left(\tilde{\tau}_{i,-3}^{BJS}, \ldots, \tilde{\tau}_{i,5}^{BJS}\right),$$

where each $\tilde{\tau}_{i,h}^{BJS}$ is the normalized version $\hat{\tau}_{i,h}^{BJS}$.

Individual-level heterogeneity

To visualize the variation in $\tilde{\tau}_{i,h}^{BJS}$, we conduct an exploratory analysis of these objects. This analysis intentionally ignores the measurement and estimation errors we discussed in Sections 2.1 and 2.4. We start by plotting the marginal distributions of $\tilde{\tau}_{i,h}^{BJS}$ (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 h.

Figure Ia focuses on marginal distributions for each h and therefore cannot address the persistence of CP and its trajectory. To investigate the persistence further, we employ a K-means algorithm (Lloyd, 1982) applied to the vector of estimated child penalties $\tilde{\boldsymbol{\tau}}_i^{BJS}$ and categorize all individuals of the same gender into three groups. The findings of this analysis are presented in Figure Ib and reveal significant variation in the CP trajectories. Notably, we observe that approximately two-thirds of the male population experience non-existent CP (orange line).

[15] See, for example, Bensnes et al. (2023) and Lundborg et al. (2024).

Simultaneously, a quarter of the sample exhibits large negative trajectories (green), while another 10% of the population demonstrates a positive trajectory relative to their counterfactual earnings growth without parenthood (blue). The situation for females is qualitatively similar; however, the trajectories differ markedly, with 28% of women appearing to exit the labor force (green), 55% transitioning to part-time work (orange), and 17% maintaining their pre-birth trajectory. These results indicate that traditional analyses reporting the average CP across the entire population may overlook significant individual-level variation. Nonetheless, these findings should be interpreted with caution due to measurement errors.

Validation of the measurement model

Once $\tilde{\tau}_{i,h}^{BJS}$ are available, summarizing them across various dimensions becomes easy and highlights key aspects of heterogeneity. These summaries can support the underlying measurement model or advise against its use. To investigate this in our empirical application, we project $\tilde{\tau}_{i,h}^{BJS}$ onto the age at first birth $A_i = E_i - B_i$ separately for individuals with different levels of education. Figure A.1 presents our estimated CPs by education level, gender, and age at first birth. We observe significant heterogeneity among these groups, which is interesting both empirically and as validation for our statistical model.

In particular, CPs are larger for less-educated women regarding participation and earnings margins. Furthermore, for these individuals, we do not observe significant anticipation effects across all ages, which suggests that $\tau_{i,h}$ is indeed linked to childbirth. The results differ for college-educated individuals, with women experiencing smaller CPs that tend to decrease uniformly with A_i over h. We also notice that the CPs of college-educated parents exhibit erratic behavior prior to giving birth, for individuals who have children before the age of 30. This indicates that we overlook crucial life cycle heterogeneity among college-educated individuals who have children at younger ages. Interestingly, this heterogeneity disappears once we focus on sufficiently older parents over 30. Motivated by these results, in the remainder of the analysis when we discuss college-educated parents we focus on those who became parents between ages of 31 and 34.

4 Policy analysis: Using unit-level estimates as an outcome

In this section, we continue with our CP analysis and explain how the ULES described in Section 3 can be used as an outcome of interest in policy evaluation exercises. Specifically, we illustrate this in the context of evaluating the effects of childcare supply expansion policies on the child penalty. This policy has been recently studied in several studies, finding a wide range of empirical evidence (Andresen and Nix, 2022a; Kleven et al., 2024; Rabaté and Rellstab,

2021).

4.1 Institutional background: Dutch childcare provision

The 2005 Dutch childcare expansion reform

We begin with a brief overview of the subsidized childcare system and its recent changes in the Netherlands (see Bettendorf et al. (2015) for more details).

Before the 2005 reform, access to childcare was not uniform. Factors such as whether 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 childcare providers, regardless of whether companies or municipalities subsidized them. For example, the system for center-based daycare (25% of children) was funded differently across the board. Most daycare centers received subsidies from employers and local governments. Although 24% of daycare centers were not subsidized, working parents could partially reduce costs through tax deductions. In addition to center-based care, around 25% of children attended playgroups, which offered part-time care for less than 4 hours a day. Finally, subsidized and unsubsidized (yet tax-deductible) options for out-of-school care resulted in a low 6% enrollment rate for 4-to 12-year-olds in center-based care in 2004 (Rabaté and Rellstab, 2021).

The 2005 reform brought significant change and created 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, since 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 use comprehensive information from the 5-digit sector classification, which has been available for all jobs in the Netherlands since 2001. These data allow us to identify childcare workers and their job locations.[16] Our childcare supply index (CCI) for each municipality is calculated by dividing the number of childcare jobs in a given municipality g and year t ($N_{g,t}^{jobs}$) by the number of children under 5 years of age in the same locality ($N_{g,t}^{children}$):

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

[16] Rabaté and Rellstab (2021) Appendix C.2 shows a strong correlation between the trends in childcare employment and the aggregate public spending on childcare, validating this index.

Figure IIa presents the variation in childcare supply per preschool-aged child across municipalities from 1999 to 2016. Each dot represents the mean CCI in a given year, whereas the shaded area represents the distribution of CCI across municipalities. The data reveal a significant range in the ratio of childcare workers to preschool children, with values ranging from 0 to 0.3.[17] The figure illustrates the substantial variation in childcare availability between different municipalities and the large increase due to the 2005 childcare expansion reform.

Remark 4.1. In the data, individuals migrate across municipalities; thus, there is no fixed association between individual i and group g. Instead, the relationship is period-specific; i.e., we can associate municipality $g(i,t)$ with individual i in period t. This structure can be accommodated within the model from Section 2.3 by defining $g(i)$ as a vector of locations the individual resides in throughout the observation period. However, this leads to a dramatic expansion of the number of possible groups and renders this unrestricted approach infeasible in practice. To simplify the analysis, we opt for a simpler solution and define $g(i)$ as $g(i, E_i)$, i.e., assign the individual to the municipality in which this person resided at the year of their first childbirth. We then focus on the level of childcare availability in location $g(i)$ in two key decision points: 1 year before first birth ($CCI_{g(i),E_i-1}$) and the contemporaneous year ($CCI_{g(i),E_i+h}$). An alternative solution would be to consider an AKM-type structure (Abowd et al., 1999), allowing for $\lambda_{g(i,t),t}$.

4.2 Descriptive analysis: CP and childcare provision levels

This section explores the descriptive relationship between childcare provision levels and CPs (Figure III). We aggregate estimated ULES 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 split the estimation between men in blue and women in orange and present the results across horizons $h \in \{0, \ldots, 5\}$. Figure IIIa presents correlations using CP estimates for earnings. Similarly, Figure IIIb presents correlations using CP estimates for participation.

For women, childcare provision levels are related to lower CP in participation and earnings. The relationship is stronger as we get further from the year of birth. This is consistent with the idea that childcare provision helps mothers return to the labor force after they finish their maternity leave, and thus shrinks the CP. Fathers exhibit a similar but weaker relationship on the earnings margin.

[17] Measurement errors might influence some of the extreme values since the job location is tied to the firm's location and may not always align with the actual municipality where the job is performed.

A naive interpretation of these results would conclude that childcare provision reduces the CP. However, this would ignore the supply-side response to a differential demand. In other words, households with higher employment and earnings potential would be more willing to pay for childcare services, resulting in higher availability in their municipality. Therefore, we shift our analysis to directly explore the 2005 Dutch childcare expansion policy.

4.3 The effect of the Dutch childcare expansion policy on the CP

Our policy analysis uses the framework described in Section 2.3, in which we use the estimated $\tilde{\tau}_{i,h}$ in Section 3.2 as an outcome. The key assumption the analysis relies on is that the policy does not directly affect the timing of childbirth. To justify this assumption, we conduct a duration analysis by relating the conditional probability of having a child to baseline levels of the policy variable, adjusting for municipality and age. Appendix B.6 discusses the underlying assumptions behind this analysis. The results are reported in Table A.1 and do not show any significant effect of the policy, which makes us more comfortable with the first assumption. To validate the rest of the assumptions, we conduct a placebo analysis and use $\tilde{\tau}^{BJS}_{i,h}$ for $h \in \{-3, -2, -1\}$ as outcomes and relate them to $(W_{g(i),E_i-3}, \ldots, W_{g(i),E_i-1})$. Results are presented in Tables A.2 and A.3 and show no relationship, thus validating our policy evaluation model.

Our primary analysis relies on specification (6) and its generalization, which accommodates multiple periods and covariates and is discussed in Appendix B.4. In particular, we relate the level of the policy in the period before childbirth, $CCI_{g(i),E_i-1}$, and the levels of the contemporaneous policy, $CCI_{g(i),E_i+h}$, to $\tilde{\tau}^{BJS}_{i,h}$ for $h \in \{0, \ldots, 5\}$. We control for the municipality of residence at the time of the event (first birth), the event year, and the age at first childbirth fixed effects. Our results in Figure IV suggest substantial heterogeneity in treatment effects, whereby more educated parents experience a greater increase in earnings due to enhanced childcare availability.[18] These heterogeneous treatment effects may provide insights into the economic mechanism underlying the relationship between childcare provision and labor supply. Specifically, we can interpret these heterogeneous treatment effects through the lens of selection: Households sort into the childcare supply they prefer. For instance, households with low labor-force attachment may prefer to raise their children during their early years and thus may sort into areas with limited childcare availability. Alternatively, these heterogeneous treatment effects can be understood from a structural perspective.

[18]This pattern also holds for the participation margin and persists when we extend our analysis to lower education groups (see Figures A.3 and A.4).

Structural interpretation of childcare expansion on parental labor supply

Using a simple household model, we illustrate a potential economic intuition behind the heterogeneous relationship between the child penalty and the contemporaneous childcare provision by gender (see Appendix C for details). When outsourcing childcare is inexpensive relative to wages, it becomes appealing for parents to outsource it, thereby increasing the likelihood of them working more and boosting household income. However, the model highlights how this intuition only emerges when the contemporaneous childcare costs fall below a certain threshold. Above that threshold, childcare becomes prohibitively expensive, and often leaves one parent (usually the mother) at home, at least during the first few years.

Furthermore, a more nuanced economic interpretation of the results is the job flexibility of parents, which may vary by education level and timing of the childcare expansion. The interplay between job type and education further highlights these dynamics. Goldin (2014) classifies jobs into two distinct types based on their wage schedules: linear and nonlinear. Linear jobs show a proportional relationship between hours worked and wages, allowing for greater flexibility in scheduling. Nonlinear jobs, in contrast, feature convex wage schedules, imposing significant penalties for reducing working hours.[19] Since nonlinear jobs typically offer higher wages, highly educated women are more inclined to pursue them, while lower-educated women tend to remain in linear jobs. As a result, childcare interventions disproportionately benefit highly educated women, since they are more likely to have jobs with demanding work-hour requirements, thereby amplifying the impact of childcare availability on their labor market outcomes. This job flexibility logic might be more relevant to contemporaneous changes in childcare, as opposed to changes to childcare during pregnancy, which have more long-term consequences, explaining the different patterns between Figures IVa and IVb.

4.4 Comparison with conventional methods

In the previous section, we discussed the benefits of using ULES estimates as objects of interest for policy evaluation. We illustrated this in the context of assessing the effects of childcare expansion on the child penalty (CP). This section will compare our method with the common one-step approach, which does not separate the measurement and policy evaluation steps, and instead implements both in one extended difference-in-differences (DiD) regression. As a result, it tends to binarize the treatment status and time variation, potentially overlooking significant data variation. Our two-step approach, in contrast, leverages the broader (nonbi-

[19] Erosa et al. (2022) formalize the trade-off between flexibility and wages in the U.S. labor market. Their model emphasizes how home production requirements within households drive women to disproportionately choose linear jobs. Yanagimoto (2024) develops a similar model to examine the role of childcare availability in shaping occupational choices. His analysis reveals that access to childcare allows women to transition from linear to nonlinear jobs by alleviating domestic labor constraints.

nary) variation in the data to provide a more nuanced understanding of policy impact, which allows us to uncover the heterogeneous treatment effects discussed in Section 4.3. This not only enriches the analysis but also offers clearer, more precise, and actionable information for policymakers and stakeholders. Finally, we will demonstrate below that the two approaches lead to qualitatively different conclusions.

One-step approach

The common one-step approach employs a DiD method that typically discretizes the pre-treatment and post-treatment periods along with the treatment status. In the context of the 2005 Dutch childcare expansion reform, pre-treatment period outcomes are defined between 2000 to 2005 and consider, for each individual, a balanced panel of time relative to first birth ($h = -1, \ldots, 5$). Similarly, post-treatment outcomes are observed between 2011 and 2016.[20]

Because the policy was implemented nationwide simultaneously, it is challenging to define a binary treatment status, a common issue encountered when evaluating any nationwide policy. The literature analyzing the effects of the expansion of childcare on the CP typically defines treatment as municipalities that experienced a CCI expansion above a certain threshold (Kleven et al., 2024; Lim and Duletzki, 2023; Rabaté and Rellstab, 2021). We adopt this convention and define treatment as municipalities with an expansion of at least ten percentage points in CCI between the pre- and post-periods:

$$T_{g(i)} \equiv \mathbf{1}\{\overline{CCI}_{g(i)}^{2010-2015} - \overline{CCI}_{g(i)}^{1999-2004} > 0.1\},$$

where we assign individual i to location $g(i)$ based on the municipality they resided in prior to childbirth, $E_i - 1$.

This variation used in the DiD specification is illustrated in Figure IIb, which plots the CCI index over time according to treatment status. We see that both the treatment and control groups show parallel trends in the pre-period. In the post-period, the treatment diverges from the control group (mechanically). Notably, the control group also responds to the policy, which is expected since the policy was implemented nationwide. This highlights a drawback of this common approach: It leaves substantial variation on the table and may lead to different conclusions. However, the main advantage of this approach is the ability to run it in a single

[20]Note that we adjusted our sample criteria to include parents with their first childbirth after 1995 to allow for a sufficient labor market horizon (e.g., to observe the horizon $h = 5$ at 2000). Furthermore, since the one-step approach does not typically segment the analysis by education, we remove the age at first birth criteria by education group, instead including all parents whose age at first birth is between 24 and 34, regardless of their education level (high school, vocational, or bachelors).

regression as follows:

$$
\begin{aligned}
Y_{i,t} = \lambda_t + \gamma_{t-B_i} + \sum_{h\neq -1} \alpha_h \mathbf{1}\left\{t - E_i = h\right\} + \sum_{h\neq -1} \rho_h \mathbf{1}\left\{t - E_i = h\right\} T_{g(i)} \\
+ \sum_{h\neq -1} \delta_h \mathbf{1}\left\{t - E_i = h\right\} \mathbf{1}\{E_i > 2005\} \\
+ \sum_{h\neq -1} \beta_h \mathbf{1}\left\{t - E_i = h\right\} T_{g(i)} \mathbf{1}\{E_i > 2005\} + v_{i,t}.
\end{aligned} \tag{9}
$$

where B_i is the birth year of individual i, which makes γ_{t-B_i} the age fixed effects, and the rest of the variables are the same as we define above.

Comparing results

Figure V presents results from estimating Equation (9). The left panel presents the results for earnings and the right panel the results for participation. We can notice that for both earnings and participation margins, the one-step approach concludes that a childcare expansion resulted in a *decrease* in earnings for both parents and participation for the mothers – a result that is hard to justify with standard economic models. These results starkly contrast with the positive effect of childcare expansion on the parental labor supply under the two-step procedure we discussed in Section 4.3.

What might explain the qualitative differences in the results of the two approaches? Significant distinctions exist in both the measurement and policy analysis. Implicitly, the one-step approach also has a measurement built-in: Regression (9) estimates the unit-level CPs and aggregates them before conducting the policy analysis. Specifically, in a given period, it implicitly compares the labor market outcomes of a parent with those of an individual of the same age who will have a child in the subsequent period.[21] In contrast, our two-step approach relies on the DiD-type comparisons outlined in Section 2.1. Because the comparisons in the one-step approach in regression (9) are cross-sectional, they can accommodate a larger group of individuals, since there is no need to adjust for pre-event outcomes. However, the downside is that it requires strong selection assumptions; for instance, model (1) suggests that unit-level fixed effects can confound such comparisons, limiting the ability of the one-step approach to control for unobserved heterogeneity. After constructing unit-level estimates, regression (9) implicitly aggregates them to the level of policy variation; namely, in four groups. Conceptually, this exercise suffers from the contamination bias explained by Goldsmith-Pinkham et al. (2024), although the magnitude of this bias can be small in some applications.

[21] Technically, additional comparisons are implicitly validated by (9), such as DiD-type comparisons of individuals from different birth cohorts across various periods (ensuring they share the same age). These questionable comparisons arise from the two-way model and are eliminated when we include $t \times A_i$ fixed effects.

The second key difference lies in the variation used in the policy evaluation exercise itself. The two-step approach uses detailed information on the policy level and its variation over time, while the one-step approach separates this variation into four groups (by treatment- and post-status). This exercise may appear cleaner because it relies on simple comparisons and does not limit any dynamic effects of the policy. However, in this particular case, such an interpretation is superficial, because the underlying policy variation dynamics *are* complex: There is no perfect pre-policy period or ideal control group of municipalities. By suggesting that it resembles the standard DiD setup, researchers do not achieve greater transparency and potentially lose significant statistical power, along with the ability to conduct a more nuanced and economically relevant analysis. Indeed, our empirical results in Figure IV suggest such dynamic considerations are important.

The methodology used in this section demonstrates the benefits of ULES measurements as outcomes in policy evaluations. 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 Conclusion

This paper introduces a new method for analyzing how units react to events, such as having a child or facing a policy change. We argue that examining these responses on a unit basis, rather than merely focusing on averages, can reveal important insights. Our approach consists of two steps: First, we *measure* how each unit is impacted by an event by comparing them with a comparable group of people who have not experienced the event. Second, we use these unit-level estimates to conduct our *causal analysis* step. We also provide an online tutorial to ease adoption of the two-step approach.[22]

We demonstrate that this two-step approach is especially effective for examining the effects of policies that change over time and exploring heterogeneous treatment effects among different groups. In our analysis, we show that the expansion of childcare in the Netherlands had varying impacts on mothers and fathers, and that these impacts depended on the parents' education level and the timing of the childcare expansion. These findings emphasize the significance of considering individual circumstances when evaluating policies.

More broadly, our analysis is relevant for empirical issues in which researchers can con-

[22] https://kazuyanagimoto.com/unitdid/

ceptually distinguish between the measurement phase and the causal analysis. For example, some commonly used relevant applications include: teacher value-added, mass layoff events, and firm wage premium estimation. Our theoretical findings are directly applicable to models in which unbiased estimation is possible. It can also be expanded to encompass more general, nonlinear models, potentially by combining it with Empirical Bayes methods.

References

Abowd, John M., Francis Kramarz, and David N. Margolis. 1999. "High Wage Workers and High Wage Firms." *Econometrica* 67 (2): 251–333. 10.1111/1468-0262.00020, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00020.

Adams, Abi, Mathias Fjællegaard Jensen, and Barbara Petrongolo. 2024. "Birth Timing and Spacing: Implications for Parental Leave Dynamics and Child Penalties."

Andresen, Martin Eckhoff, and Emily Nix. 2022a. "Can the child penalty be reduced? Evaluating multiple policy interventions." http://hdl.handle.net/10419/268059, Issue: 983 tex.copyright: http://www.econstor.eu/dspace/Nutzungsbedingungen.

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.

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, Publisher: The University of Chicago Press.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, . 10.2307/j.ctvcm4j72.

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.

Arellano, Manuel, and Olympia Bover. 1995. "Another look at the instrumental variable estimation of error-components models." *Journal of Econometrics* 68 (1): 29–51. 10.1016/0304-4076(94)01642-D.

Arkhangelsky, Dmitry, and Guido Imbens. 2024. "Causal models for longitudinal and panel data: a survey." *The Econometrics Journal* 27 (3): C1–C61. 10.1093/ectj/utae014.

Arkhangelsky, Dmitry, and Guido W Imbens. 2022. "Doubly robust identification for causal panel data models." *The Econometrics Journal* 25 (3): 649–674. 10.1093/ectj/utac019.

Arkhangelsky, Dmitry, Guido W. Imbens, Lihua Lei, and Xiaoman Luo. 2024a. "Design-robust two-way-fixed-effects regression for panel data." *Quantitative Economics* 15 (4): 999–1034. 10.3982/QE1962, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE1962.

Arkhangelsky, Dmitry, and Aleksei Samkov. 2024. "Sequential Synthetic Difference in Differences." March. 10.48550/arXiv.2404.00164, arXiv:2404.00164 [econ].

Arkhangelsky, Dmitry, Kazuharu Yanagimoto, and Tom Zohar. 2024b. "Flexible Analysis of Individual Heterogeneity in Event Studies: Application to the Child Penalty." March. 10.48550/arXiv.2403.19563, arXiv:2403.19563 [econ].

Athey, Susan, and Stefan Wager. 2021. "Policy Learning With Observational Data." *Econometrica* 89 (1): 133–161. 10.3982/ECTA15732, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA15732.

Bensnes, Simon, Ingrid Huitfeldt, and Edwin Leuven. 2023. "Reconciling Estimates of the Long-Term Earnings Effect of Fertility." *SSRN Electronic Journal*. 10.2139/ssrn.4464587.

Bettendorf, Leon J. H., Egbert L. W. Jongen, and Paul Muller. 2015. "Childcare subsidies and labour supply — Evidence from a large Dutch reform." *Labour Economics* 36 112–123. 10.1016/j.labeco.2015.03.007.

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.

Borusyak, Kirill, and Peter Hull. 2023. "Nonrandom Exposure to Exogenous Shocks." *Econometrica* 91 (6): 2155–2185. 10.3982/ECTA19367, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA19367.

Borusyak, Kirill, and Peter Hull. 2024. "Negative Weights Are No Concern in Design-Based Specifications." *AEA Papers and Proceedings* 114 597–600. 10.1257/pandp.20241046.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel. 2024a. "Design-based identification with formula instruments: A review." *The Econometrics Journal* utae003. 10.1093/ectj/utae003.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024b. "Revisiting Event-Study Designs: Robust and Efficient Estimation." *The Review of Economic Studies* rdae007. 10.1093/restud/rdae007.

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.

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, Publisher: [Wiley, Econometric Society].

Chernozhukov, V, W K Newey, and R Singh. 2023. "A simple and general debiased machine learning theorem with finite-sample guarantees." *Biometrika* 110 (1): 257–264. 10.1093/biomet/asac033.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. 2017. "Central limit theorems and bootstrap in high dimensions." *The Annals of Probability* 45 (4): 2309–2352. 10.1214/16-AOP1113, Publisher: Institute of Mathematical Statistics.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2018. "Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India." June. 10.3386/w24678.

Chernozhukov, Victor, Whitney K. Newey, and Rahul Singh. 2022. "Automatic Debiased Machine Learning of Causal and Structural Effects." *Econometrica* 90 (3): 967–1027. 10.3982/ECTA18515, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA18515.

De Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2020. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." *SSRN Electronic Journal*. 10.2139/ssrn.3731856.

Erosa, Andrés, Luisa Fuster, Gueorgui Kambourov, and Richard Rogerson. 2022. "Hours, Occupations, and Gender Differences in Labor Market Outcomes." *American Economic Journal: Macroeconomics* 14 (3): 543–590. 10.1257/mac.20200318.

Foster, Dylan J., and Vasilis Syrgkanis. 2023. "Orthogonal statistical learning." *The Annals of Statistics* 51 (3): 879–908. 10.1214/23-AOS2258, Publisher: Institute of Mathematical Statistics.

Freyaldenhoven, Simon, Christian Hansen, Jorge Pérez Pérez, and Jesse M. Shapiro. 2021. "Visualization, Identification, and Estimation in the Linear Panel Event-Study Design." August. 10.3386/w29170.

Gallen, Yana. 2019. "The effect of parental leave extensions on firms and coworkers."

Goldin, Claudia. 2014. "A Grand Gender Convergence: Its Last Chapter." *American Economic Review* 104 (4): 1091–1119. 10.1257/aer.104.4.1091.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2024. "Contamination Bias in Linear Regressions." *American Economic Review* 114 (12): 4015–4051. 10.1257/aer.20221116.

Graham, Bryan S., and James L. Powell. 2012. "Identification and Estimation of Average Partial Effects in "Irregular" Correlated Random Coefficient Panel Data Models." *Econometrica* 80 (5): 2105–2152. 10.3982/ECTA8220, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA8220.

Holtz-Eakin, Douglas, Whitney Newey, and Harvey S. Rosen. 1988. "Estimating Vector Autoregressions with Panel Data." *Econometrica* 56 (6): 1371–1395. 10.2307/1913103, Publisher: [Wiley, Econometric Society].

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press, . 10.1017/CBO9781139025751.

Karademir, Sencer, Jean-William P. Laliberté, and Stefan Staubli. 2024. "The Multigenerational Impact of Children and Childcare Policies." March. 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, Publisher: Institute of Mathematical Statistics and Bernoulli Society.

Kleven, Henrik, Camille Landais, Johanna Posch, Andreas Steinhauer, and Josef Zweimüller. 2024. "Do Family Policies Reduce Gender Inequality? Evidence from 60 Years of Policy Experimentation." *American Economic Journal: Economic Policy* 16 (2): 110–149. 10.1257/pol.20210346.

Kleven, Henrik, Camille Landais, and Jakob Egholt Søgaard. 2019. "Children and Gender Inequality: Evidence from Denmark." *American Economic Journal: Applied Economics* 11 (4): 181–209. 10.1257/app.20180010.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *The Review of Economic Studies* 76 (3): 1071–1102, https://econpapers.repec.org/article/ouprestud/v_3a76_3ay_3a2009_3ai_3a3_3ap_3a1071-1102.htm, Publisher: Review of Economic Studies Ltd.

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, Conference Name: IEEE Transactions on Information Theory.

Lundborg, Petter, Erik Plug, and Astrid Würtz Rasmussen. 2024. "Is There Really a Child Penalty in the Long Run? New Evidence from IVF Treatments." *SSRN Electronic Journal*. 10.2139/ssrn.4813455.

Muris, Chris, and Konstantin Wacker. 2022. "Estimating interaction effects with panel data." November. 10.48550/arXiv.2211.01557, arXiv:2211.01557 [econ].

Nie, X, and S Wager. 2021. "Quasi-oracle estimation of heterogeneous treatment effects." *Biometrika* 108 (2): 299–319. 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, Publisher: CPB Netherlands Bureau for Economic Policy Analysis Version Number: CPB discussion paper, 424.

Robinson, P. M. 1988. "Root-N-Consistent Semiparametric Regression." *Econometrica* 56 (4): 931–954. 10.2307/1912705, Publisher: [Wiley, Econometric Society].

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. 10.1093/ectj/utaa027.

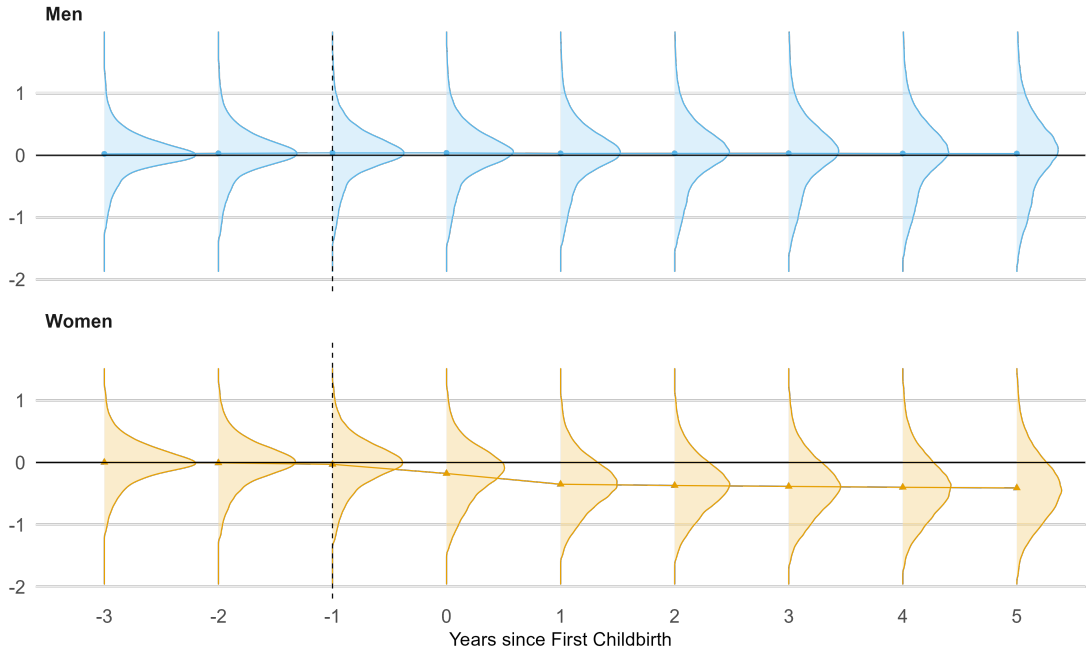
Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics* 225 (2): 175–199. 10.1016/j.jeconom.2020.09.006.

Yanagimoto, Kazuharu. 2024. "Why not Choose a Better Job? Flexibility, Social Norms, and Gender Gaps in Japan." March, https://www.cemfi.es/ftp/wp/2405.pdf.

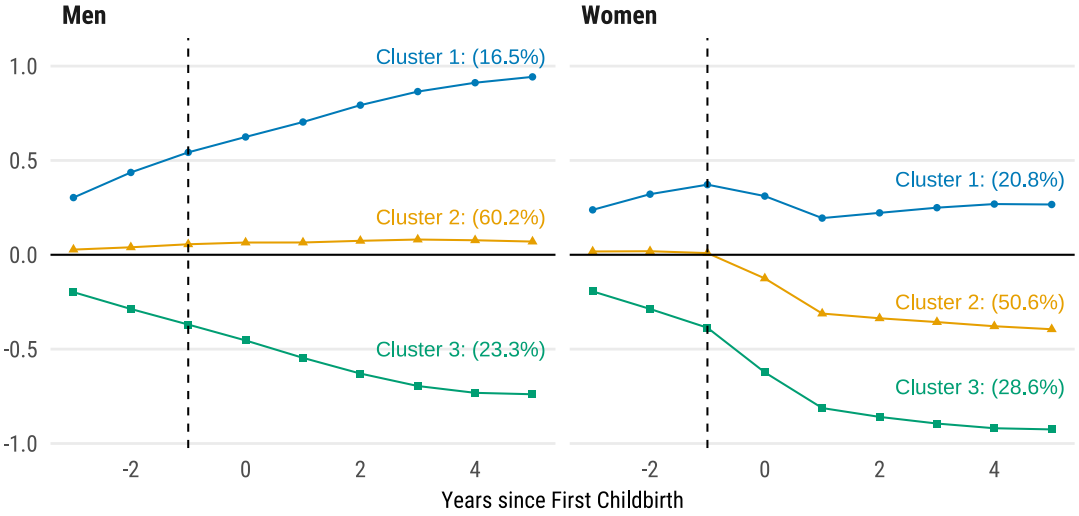
Figures

Figure I: Heterogeneity in child penalties (CP)

(a) Distribution of Individual CP



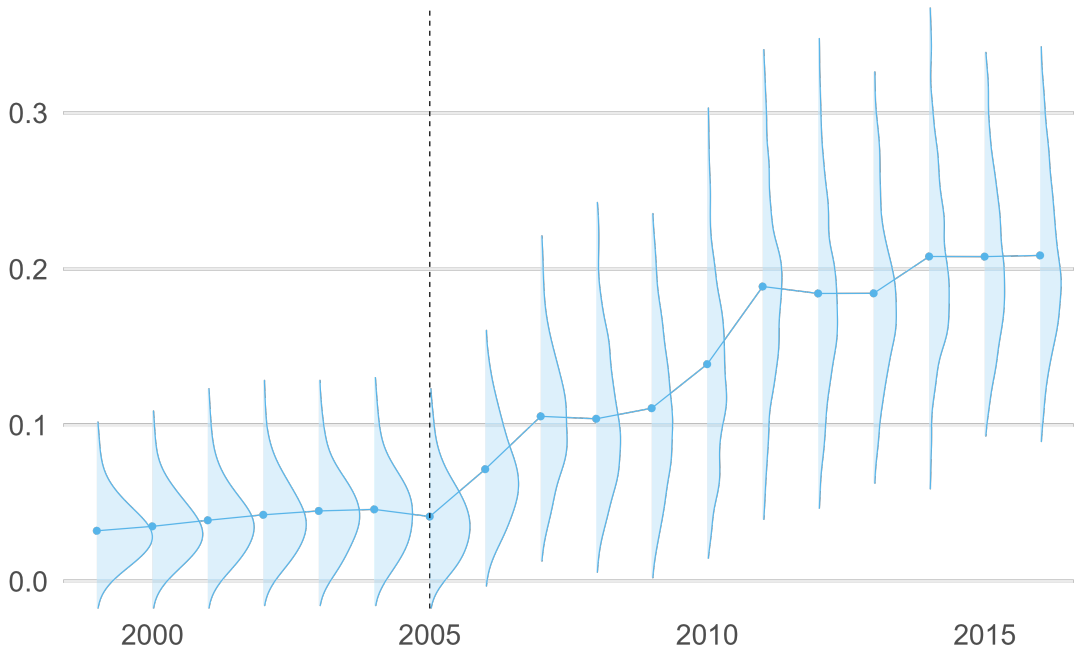
(b) Different CP Paths (K-Means)



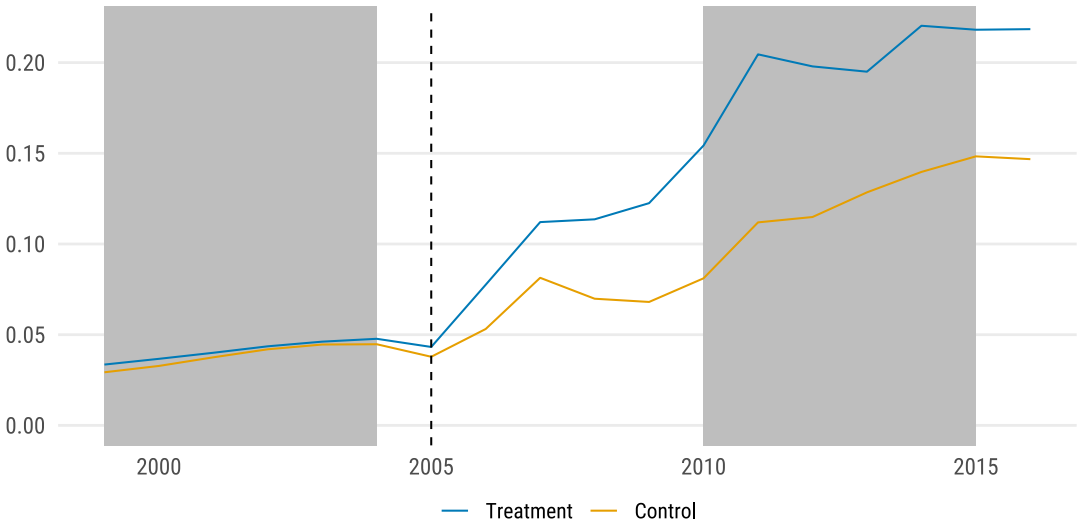
Notes: These figures present the variation in child penalty (CP) estimates. Figure Ia plots the marginal distributions of $\tilde{\tau}_{i,h}$ as described in Section 3.2 for earnings, by time relative to first childbirth and genders, pooled across all birth cohorts and education. The dots represent the mean of each distribution. Figure Ib shows the results from applying a K-means algorithm to the vector of estimated child penalties $\tilde{\boldsymbol{\tau}}_i$ and classifying all individuals of the same gender into three groups.

Figure II: Childcare supply expansion

(a) Distribution of childcare index by municipality



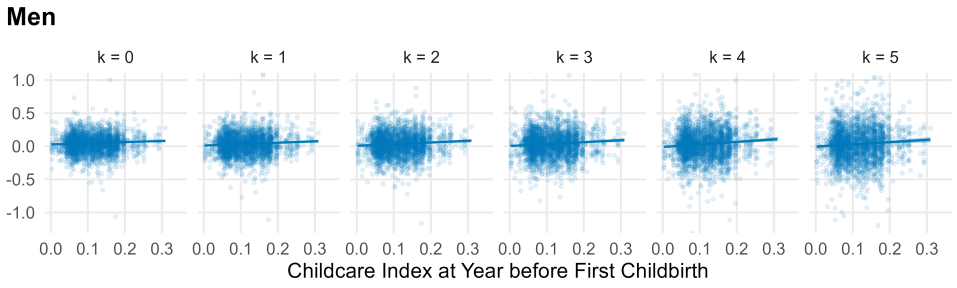
(b) Simplified 2×2 DiD design

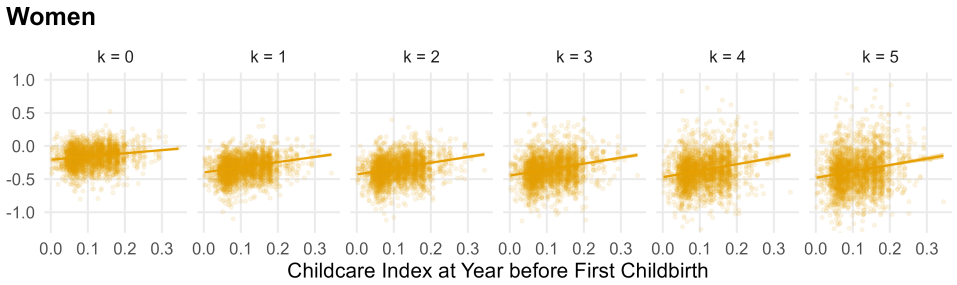


Notes: These figures present the variation in childcare supply per preschool-aged children across municipalities from 1999 to 2016.Our childcare supply index (CCI) for each municipality is calculated by dividing the number of childcare jobs in a given municipality m and year t ($N_{m,t}^{jobs}$) by the number of children under 5 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 childcare expansion reform. Dots represent the mean CCI in a given year, whereas the shaded area represents the distribution of CCI across municipalities in that year. Panel (b) illustrates the equivalent simplified 2×2 DiD design, where the time variation is binary (gray area) and treatment is binary (see Section 4.4). The pre-treatment period includes individuals at $h = -1, \ldots, 5$ in the period 2000-2005 as fully non-treated. The post-treatment period comprises individuals at $h = -1, \ldots, 5$ in the period 2011-2016 deemed fully treated. Treatment is defined as municipalities with an expansion of at least 10 percentage points in CCI between pre- and post-periods.

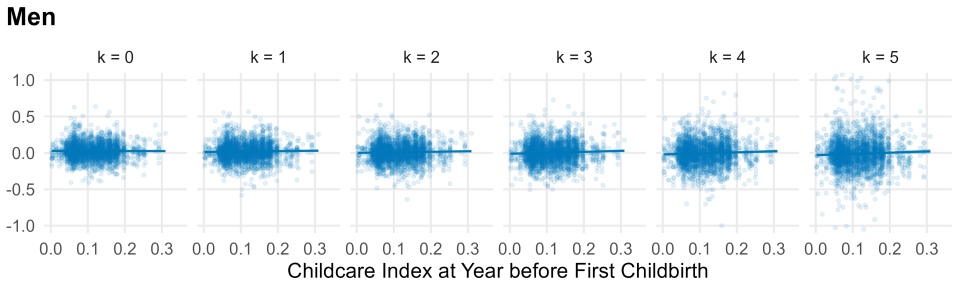
Figure III: Correlation between Childcare Provision Levels and Child Penalties

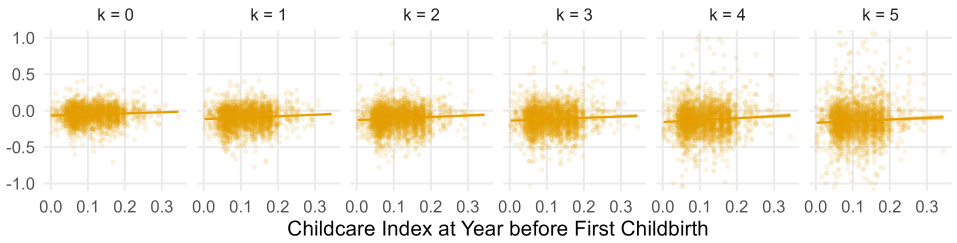
(a) Earnings





(b) Participation

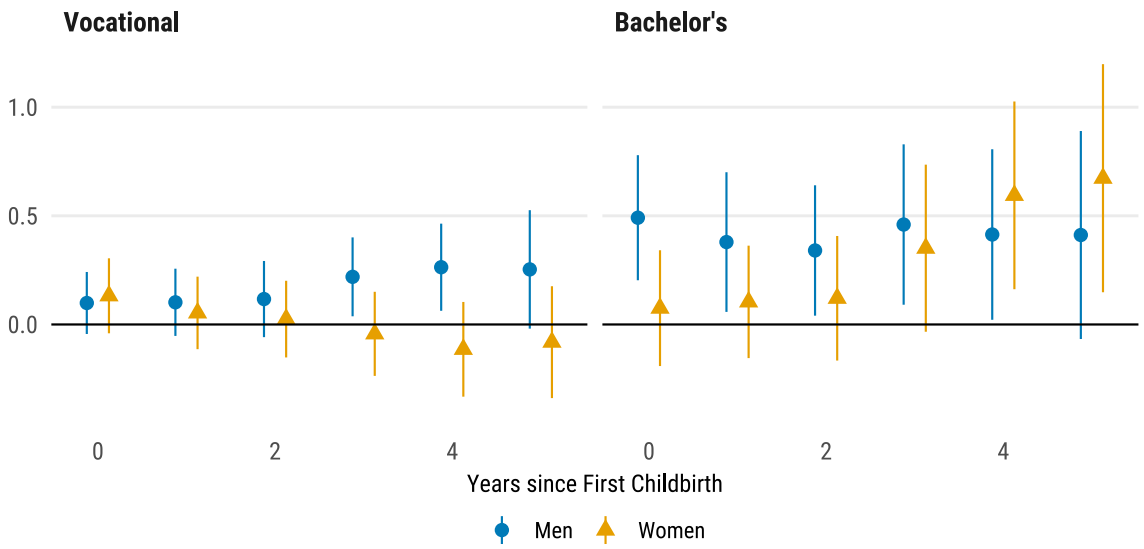




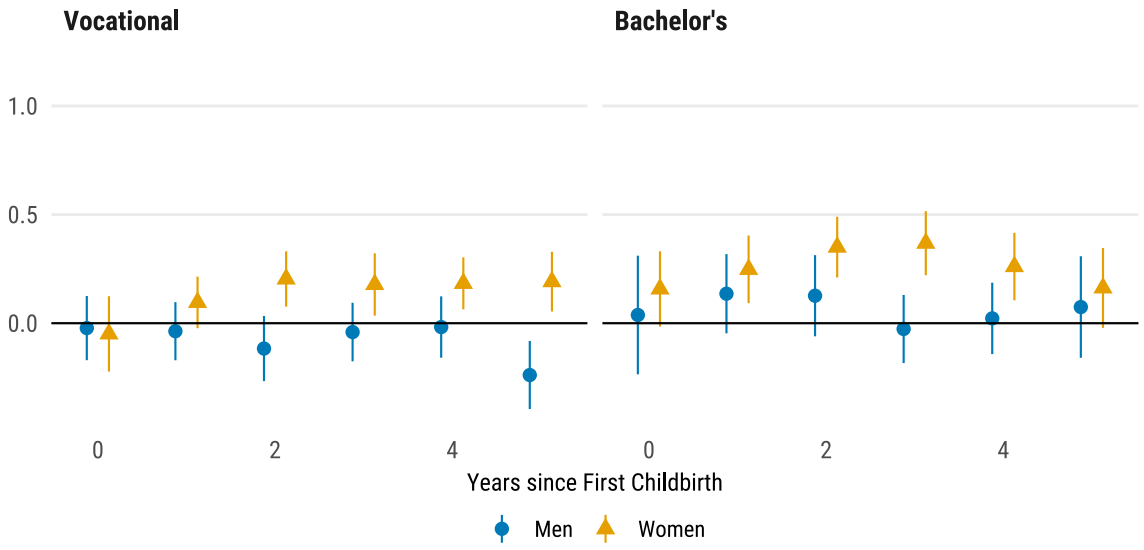
Notes: This figure presents the correlation between child penalties (CP) and the childcare provision index (CCI). We aggregate the estimated individual CP at the municipality times year-of-conception level, following the empirical strategy described in Section 3.2, and plot them against the CCI. For security reasons, only the cells with more than 10 samples are used. 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 (h). Figure IIIa presents the correlations using CP estimates for earnings. Similarly, Figure IIIb presents the correlations using CP estimates for participation. Our CCI for each municipality is calculated by dividing the number of childcare jobs in a given municipality m and year t ($N_{g,t}^{jobs}$) by the number of children under 5 years of age in the same locality ($N_{g,t}^{children}$).

Figure IV: Effect of the childcare provision expansion on child penalties of earnings

(a) $E_i - 1$

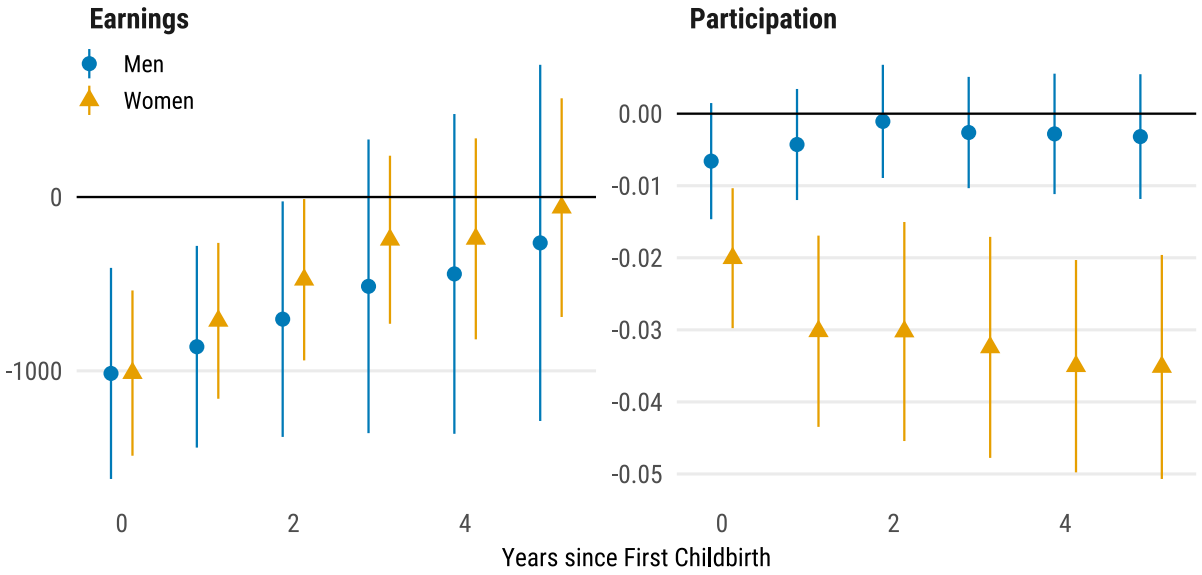


(b) $E_i + h$



Notes: This figure presents the effect of the childcare provision expansion on child penalties (CP) in earnings, split by the highest education attained. We use the specification in (6) and its generalization discussed in 4.3. In particular, we regress the individual-level child penalties $\hat{\tau}_{i,h}^{BJS}$ of earnings on the childcare index in the period before childbirth, $CCI_{g(i),E_i-1}$, and the levels of the contemporaneous policy, $CCI_{g(i),E_i+h}$, with fixed effects of the municipality at childbirth $g(i)$, event time E_i, and the age at childbirth A_i. We split the estimation between men (blue) and women (orange) and by final education attainment. IVa presents the coefficients and their 95% confidential intervals for $CCI_{g(i),E_i-1}$ and IVb for $CCI_{g(i),E_i+h}$. 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 ($N_{g,t}^{jobs}$) by the number of children under 5 years of age in the same locality ($N_{g,t}^{children}$).

Figure V: Effect of childcare expansion: one-step approach



Notes: This figure presents discrete difference-in-differences estimates of the effects of the childcare expansion on child penalties, described in Section 4.4. The pre-/post-period are 2000-2005 and 2011-2016, and the treatment group is defined as the municipality where the childcare index increased by 10 percentages points from the pre-period to the post-period. 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 ($N_{g,t}^{jobs}$) by the number of children under 5 years of age in the same locality ($N_{g,t}^{children}$)

A Figures and tables

Table A.1: Diagnostics of the childcare provision expansion policy and timing of childbirth

	$p^b_{g,t}$	$p^b_{g,t} - p^{b-1}_{g,t-1}$
	(1)	(2)
$CCI_{g,t-1}$	0.003	
	(0.008)	
$CCI_{g,t-1} - CCI_{g,t-2}$		0.004
		(0.009)
N	218,416	188,115
R^2	0.310	0.000
FE: Municipality (g) $\times$ Age ($t - b$)	X	

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Notes: The table presents the relationship between the conditional probability of having a child and the lagged levels of the childcare index. In the first column, we estimate a linear equation via OLS using the conditional probability of having a child in a given period as the outcome and the lagged level of the childcare index with the interaction of municipality and age-fixed effects as regressors. In the second column, we use the difference in conditional probabilities as the outcome and the difference in lagged policy levels as the regressor. See Appendix B.6 for the relevant discussion.

Table A.2: Placebo analysis of childcare index and pre-period child penalties (earnings)

(a) Men

	High School			Vocational			Bachelor's		
	$\tilde{\tau}_{i,-3}$	$\tilde{\tau}_{i,-2}$	$\tilde{\tau}_{i,-1}$	$\tilde{\tau}_{i,-3}$	$\tilde{\tau}_{i,-2}$	$\tilde{\tau}_{i,-1}$	$\tilde{\tau}_{i,-3}$	$\tilde{\tau}_{i,-2}$	$\tilde{\tau}_{i,-1}$
$CCI_{g(i),E_i-3}$	-0.205	-0.316	0.157	0.079	-0.036	0.008	0.088	0.065	-0.038
	(0.201)	(0.271)	(0.218)	(0.071)	(0.080)	(0.096)	(0.084)	(0.092)	(0.104)
$CCI_{g(i),E_i-2}$	0.152	0.170	0.089	-0.073	0.089	0.137	0.118	0.157	0.209
	(0.221)	(0.261)	(0.246)	(0.078)	(0.103)	(0.100)	(0.120)	(0.146)	(0.158)
$CCI_{g(i),E_i-1}$	0.002	-0.085	0.254	0.127*	0.100	0.062	0.033	0.063	0.211
	(0.174)	(0.250)	(0.205)	(0.057)	(0.080)	(0.081)	(0.137)	(0.133)	(0.139)
N	26,823	26,823	26,823	72,242	72,242	72,242	34,920	34,920	34,920
R^2	0.018	0.017	0.019	0.009	0.010	0.011	0.015	0.015	0.017
Municipality FE	X	X	X	X	X	X	X	X	X
Event year FE	X	X	X	X	X	X	X	X	X
Age at event FE	X	X	X	X	X	X	X	X	X

+ p < 0.1, * p < 0.05, ** p < 0.01

(b) Women

	High School			Vocational			Bachelor's		
	$\tilde{\tau}_{i,-3}$	$\tilde{\tau}_{i,-2}$	$\tilde{\tau}_{i,-1}$	$\tilde{\tau}_{i,-3}$	$\tilde{\tau}_{i,-2}$	$\tilde{\tau}_{i,-1}$	$\tilde{\tau}_{i,-3}$	$\tilde{\tau}_{i,-2}$	$\tilde{\tau}_{i,-1}$
$CCI_{g(i),E_i-3}$	-0.292	-0.082	-0.125	-0.025	-0.022	-0.008	0.072	-0.047	0.062
	(0.213)	(0.208)	(0.204)	(0.079)	(0.079)	(0.090)	(0.094)	(0.134)	(0.153)
$CCI_{g(i),E_i-2}$	0.203	0.064	-0.042	0.114	0.106	0.085	0.024	0.000	-0.053
	(0.236)	(0.212)	(0.209)	(0.072)	(0.085)	(0.100)	(0.101)	(0.113)	(0.139)
$CCI_{g(i),E_i-1}$	0.002	0.099	-0.019	0.018	0.036	0.068	-0.006	0.049	-0.001
	(0.149)	(0.181)	(0.216)	(0.061)	(0.076)	(0.089)	(0.103)	(0.112)	(0.137)
N	24,652	24,652	24,652	66,646	66,646	66,646	29,261	29,261	29,261
R^2	0.022	0.026	0.027	0.011	0.014	0.015	0.019	0.019	0.023
Municipality FE	X	X	X	X	X	X	X	X	X
Event year FE	X	X	X	X	X	X	X	X	X
Age at event FE	X	X	X	X	X	X	X	X	X

+ p < 0.1, * p < 0.05, ** p < 0.01

Notes: These tables present the pre-period relation between the childcare index $CCI_{g,E_i-3}, \ldots, CCI_{g,E_i-1}$ and the child penalties $\tilde{\tau}_{i,-3}, \ldots, \tilde{\tau}_{i,-1}$. We run this placebo test splitting by gender and education and with the fixed effects of municipality, event year, and age at first childbirth.

Table A.3: Placebo analysis of childcare index and pre-period child penalties (Participation)

(a) Men

	High School			Vocational			Bachelor's		
	$\tilde{\tau}_{i,-3}$	$\tilde{\tau}_{i,-2}$	$\tilde{\tau}_{i,-1}$	$\tilde{\tau}_{i,-3}$	$\tilde{\tau}_{i,-2}$	$\tilde{\tau}_{i,-1}$	$\tilde{\tau}_{i,-3}$	$\tilde{\tau}_{i,-2}$	$\tilde{\tau}_{i,-1}$
$CCI_{g(i),E_i-3}$	0.025	-0.035	0.130	0.040	-0.043	-0.054	-0.049	0.024	-0.017
	(0.128)	(0.128)	(0.133)	(0.050)	(0.058)	(0.063)	(0.058)	(0.083)	(0.076)
$CCI_{g(i),E_i-2}$	-0.021	0.129	-0.043	-0.015	0.094	0.158*	0.089	-0.015	0.080
	(0.145)	(0.178)	(0.161)	(0.052)	(0.079)	(0.063)	(0.054)	(0.081)	(0.084)
$CCI_{g(i),E_i-1}$	0.104	0.024	0.131	0.091*	0.087+	0.017	0.027	0.071	0.062
	(0.115)	(0.123)	(0.125)	(0.039)	(0.045)	(0.048)	(0.054)	(0.058)	(0.078)
N	26,823	26,823	26,823	72,242	72,242	72,242	34,920	34,920	34,920
R^2	0.017	0.019	0.019	0.008	0.008	0.009	0.014	0.017	0.018
Municipality FE	X	X	X	X	X	X	X	X	X
Event year FE	X	X	X	X	X	X	X	X	X
Age at event FE	X	X	X	X	X	X	X	X	X

+ p < 0.1, * p < 0.05, ** p < 0.01

(b) Women

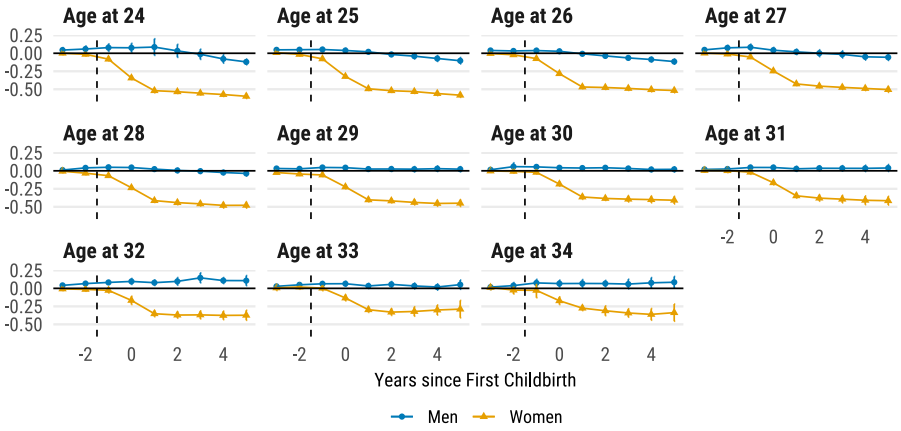
	High School			Vocational			Bachelor's		
	$\tilde{\tau}_{i,-3}$	$\tilde{\tau}_{i,-2}$	$\tilde{\tau}_{i,-1}$	$\tilde{\tau}_{i,-3}$	$\tilde{\tau}_{i,-2}$	$\tilde{\tau}_{i,-1}$	$\tilde{\tau}_{i,-3}$	$\tilde{\tau}_{i,-2}$	$\tilde{\tau}_{i,-1}$
$CCI_{g(i),E_i-3}$	0.158	0.068	0.198	0.086	0.126*	0.072	-0.013	-0.073	-0.010
	(0.134)	(0.153)	(0.138)	(0.053)	(0.052)	(0.063)	(0.081)	(0.078)	(0.092)
$CCI_{g(i),E_i-2}$	-0.035	0.157	-0.140	0.073	-0.009	0.098	-0.008	-0.020	-0.157
	(0.149)	(0.143)	(0.147)	(0.054)	(0.063)	(0.067)	(0.091)	(0.086)	(0.099)
$CCI_{g(i),E_i-1}$	0.021	0.072	0.049	-0.053	0.041	0.008	0.053	0.064	0.164
	(0.116)	(0.112)	(0.144)	(0.045)	(0.047)	(0.054)	(0.079)	(0.072)	(0.100)
N	24,652	24,652	24,652	66,646	66,646	66,646	29,261	29,261	29,261
R^2	0.017	0.019	0.019	0.007	0.009	0.011	0.017	0.020	0.024
Municipality FE	X	X	X	X	X	X	X	X	X
Event year FE	X	X	X	X	X	X	X	X	X
Age at event FE	X	X	X	X	X	X	X	X	X

+ p < 0.1, * p < 0.05, ** p < 0.01

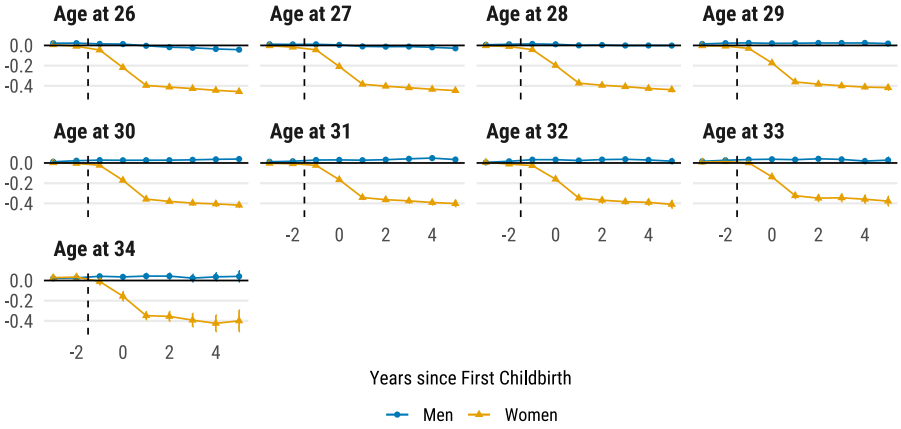
Notes: These tables present the pre-period relation between the childcare index $CCI_{g,E_i-3}, \ldots, CCI_{g,E_i-1}$ and the child penalties $\tilde{\tau}_{i,-3}, \ldots, \tilde{\tau}_{i,-1}$. We run this placebo test splitting by gender and education and with the fixed effects of municipality, event year, and age at first childbirth.

Figure A.1: Child penalties by age at first childbirth and education level (earnings)

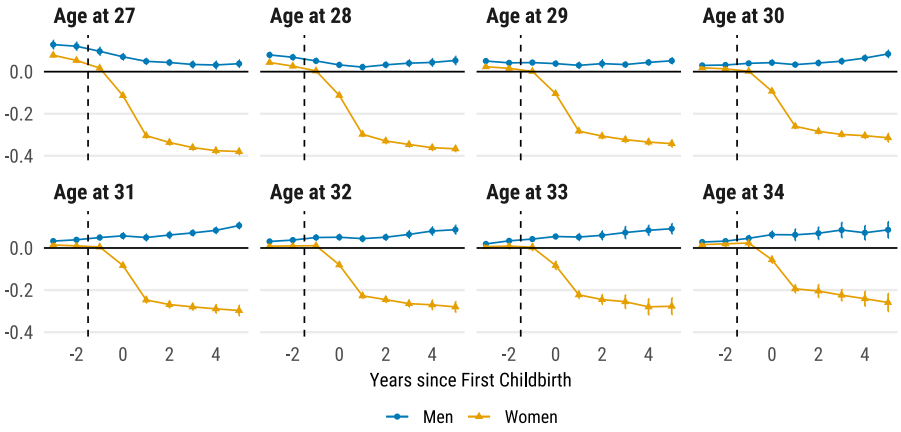
(a) High school



(b) Vocational



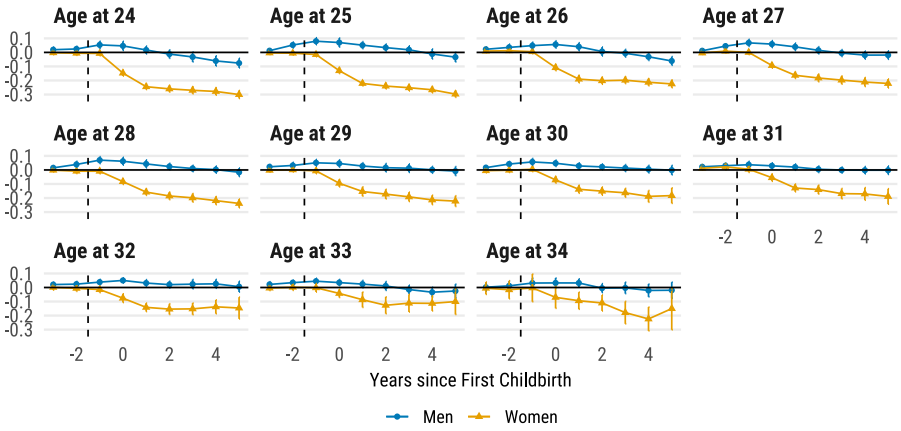
(c) Bachelor's



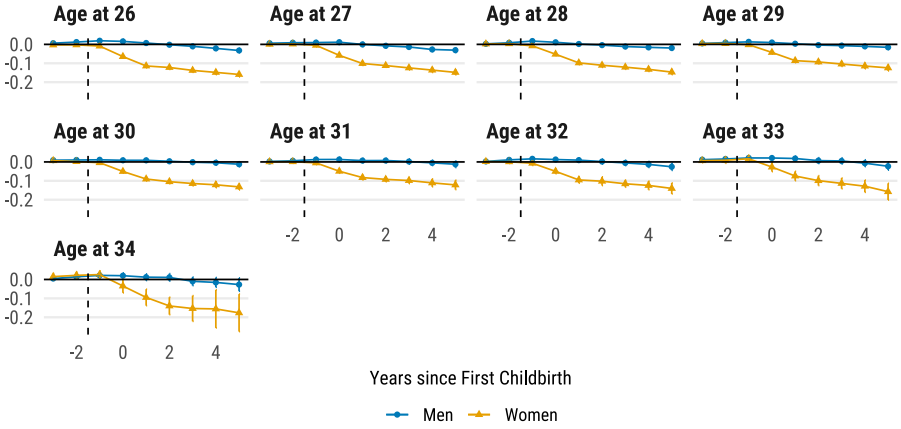
Notes: This figure presents CP estimates for yearly earnings, aggregated by observables of interest—at age of first childbirth and education level—as described in Section 3.2. Figure A.1a reports the average CP of individuals with high school diplomas as the highest obtained degree, across age at first childbirth at all horizons. Similarly, Figure A.1b reports for those who graduated from vocational school. Finally, Figure A.1c reports the results for college-educated individuals.

Figure A.2: CP by age at first childbirth and education level (participation)

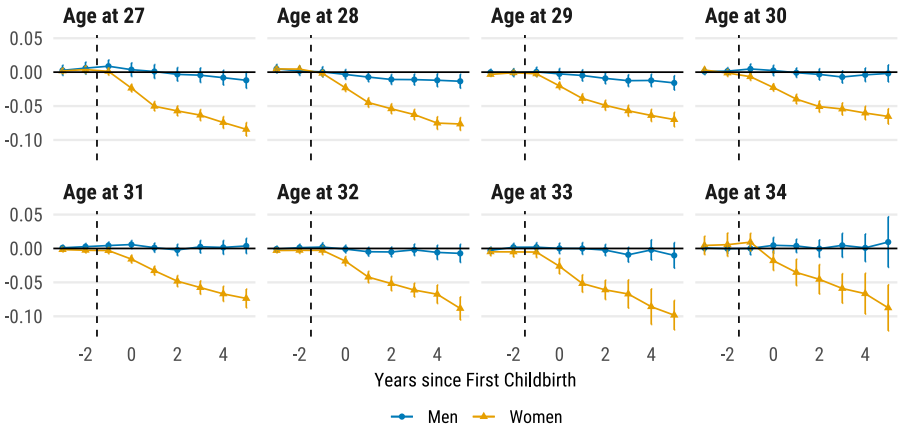
(a) High school



(b) Vocational



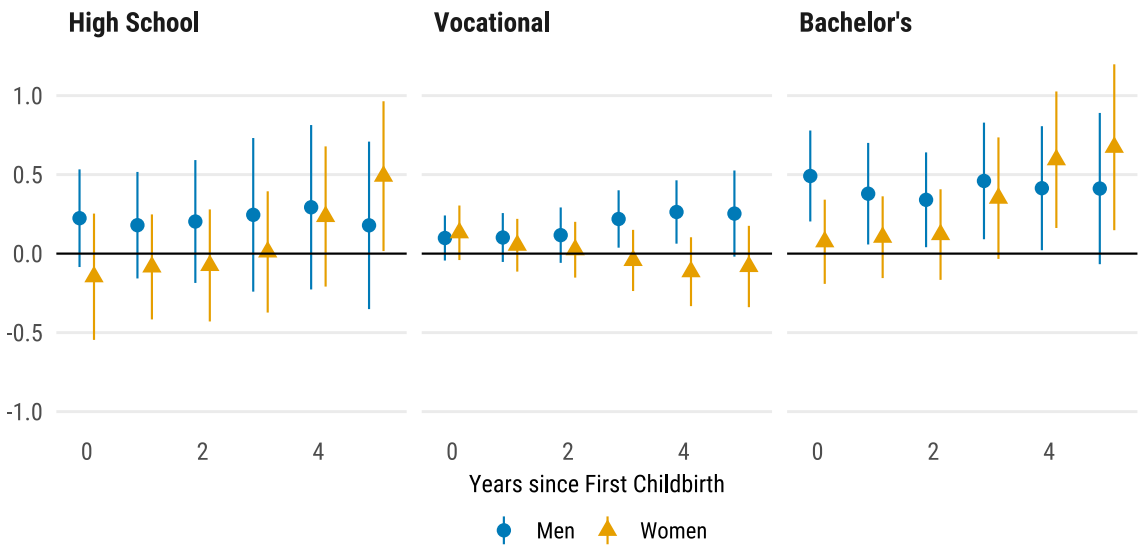
(c) Bachelor's



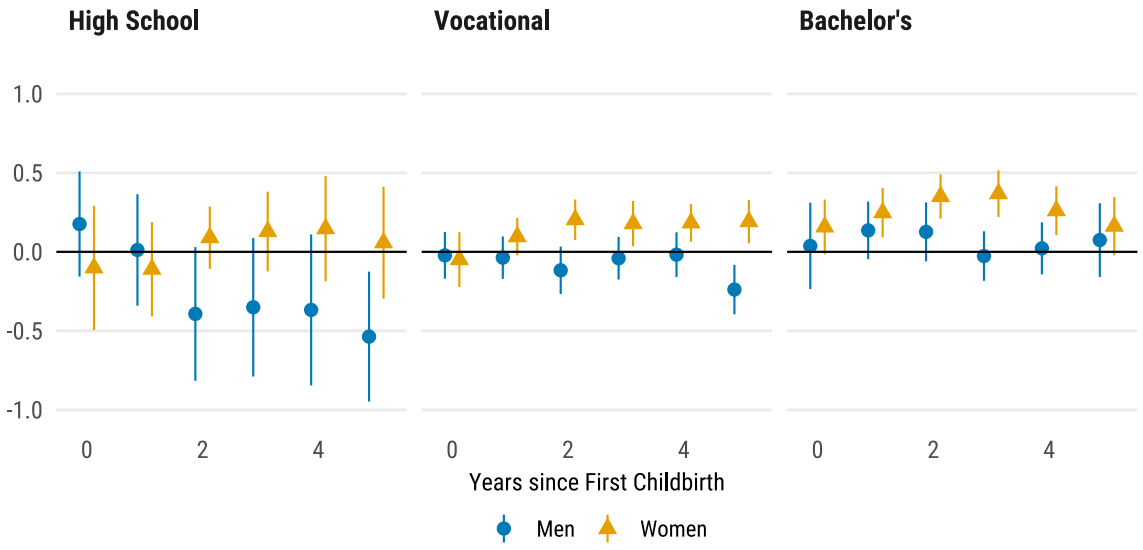
Notes: This figure presents CP estimates for labor-market participation margin, aggregated by observables of interest—age at first childbirth and education level—as described in Section 3.2. Figure A.2a reports the average CP of individuals with high school diplomas as the highest obtained degree, across age at first childbirth at all horizons. Similarly, Figure A.2b reports for those who graduated from vocational school. Finally, Figure A.2c reports the results for college-educated individuals.

Figure A.3: Effect of the childcare provision expansion on child penalties (earnings)

(a) $E_i - 1$



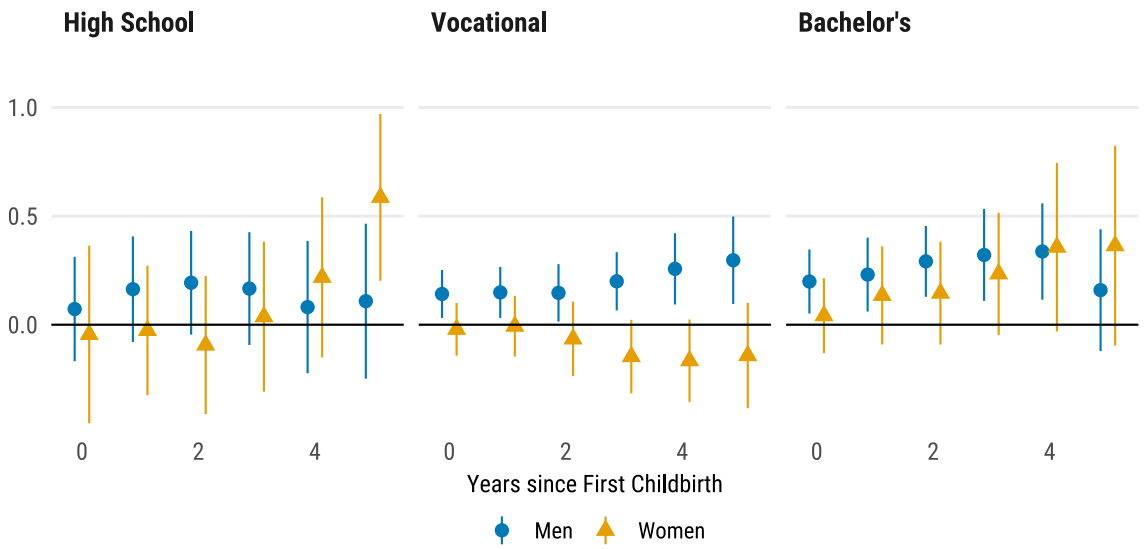
(b) $E_i + h$



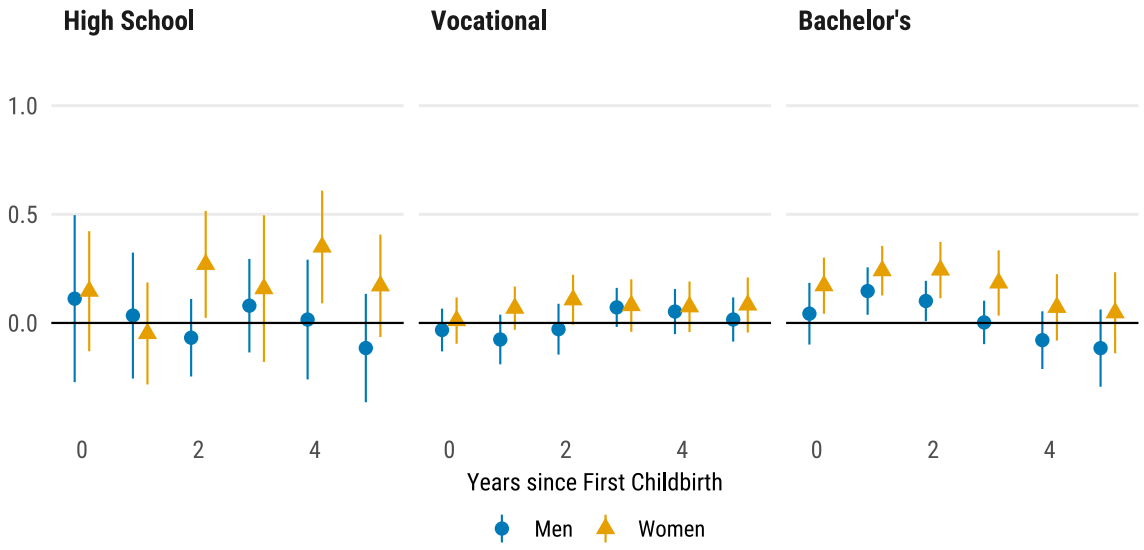
Notes: This figure presents the effect of the childcare provision expansion on child penalties (CP) in earnings, split by the highest education attained. We use the specification in (6) and its generalization discussed in 4.3. In particular, we regress individual-level child penalties $\hat{\tau}_{i,h}^{BJS}$ on the childcare index in the period before childbirth, $CCI_{g(i),E_i-1}$, and the levels of the contemporaneous policy, $CCI_{g(i),E_i+h}$, with fixed effects of municipality at childbirth $g(i)$, event time E_i, and the age at childbirth A_i. We split the estimation between men (blue) and women (orange) and by the final education attainment. A.3a presents the coefficients and their 95% confidence intervals for $CCI_{g(i),E_i-1}$ and A.3b for $CCI_{g(i),E_i+h}$. Our childcare supply index (CCI) for each municipality is calculated by dividing the number of childcare jobs in a given municipality g and year t ($N_{g,t}^{jobs}$) by the number of children under 5 years of age in the same locality ($N_{g,t}^{children}$).

Figure A.4: Effect of the childcare provision expansion on child penalties (Participation)

(a) $E_i - 1$



(b) $E_i + h$



Notes: This figure presents the effect of childcare provision expansion on child penalties (CP) in participation, split by the highest education attained. We use the specification in (6) and its generalization discussed in 4.3. In particular, we regress the individual-level child penalties $\hat{\tau}_{i,h}^{BJS}$ on the childcare index in the period before childbirth, $CCI_{g(i),E_i-1}$, and the levels of the contemporaneous policy, $CCI_{g(i),E_i+h}$, with fixed effects of municipality at childbirth $g(i)$, event time E_i, and the age at childbirth A_i. We split the estimation between men (blue) and women (orange) and by the final education attainment. A.4a presents the coefficients and their 95% confidence intervals for $CCI_{g(i),E_i-1}$ and A.4b for $CCI_{g(i),E_i+h}$. Our childcare supply index (CCI) for each municipality is calculated by dividing the number of childcare jobs in a given municipality g and year t ($N_{g,t}^{jobs}$) by the number of children under 5 years of age in the same locality ($N_{g,t}^{children}$).

B Theoretical details

B.1 Probability model

Consider a set of groups $\mathcal{G}$, where $|\mathcal{G}|$ is finite, with $g \in \mathcal{G}$ being a generic group. Consider a random element W—the observed policy intervention. For each $g \in \mathcal{G}$, let $\mathbb{P}^W_{|g}$ be the g-specific distribution of W.

Consider a unit characterized by observed attributes X with a birth cohort B being one of them. Also, each unit is characterized by a random function of W: $E^\star(\cdot)$—the potential event calendar time. The relevant potential event time is right-censored:

$$E(\cdot) := E^\star(\cdot)\mathbf{1}\{E^\star(\cdot) \leq T_{\max}\} + \infty \mathbf{1}\{E^\star(\cdot) > T_{\max}\},$$

where $T_{\max}$ is deterministic and known. Finally, each unit is characterized by potential outcomes $Y_t(\cdot)$, which are functions of W and are observed for $t \in [B + A_{\min}, B + A_{\max}]$, where $A_{\min} < A_{\max}$ are deterministic and known. We collect all these quantities into a random function of W: $D(\cdot) := (Y_{B+A_{\min}}(\cdot), \ldots, Y_{B+A_{\max}}(\cdot), E(\cdot), X)$. For each g we let $\mathbb{P}^{D(\cdot)}_{|g}$ be the g-specific distribution of $D(\cdot)$. For consistency, we assume that $B + A_{\max} \leq T_{\max}$ a.s. for each group g.

We make our first assumption about the data.

Assumption B.1. (Strict Exogeneity)
For each g define the product distribution:

$$\mathbb{P}^{\mathcal{D}}_{|g} := \mathbb{P}^W_{|g} \times \mathbb{P}^{D(\cdot)}_{|g}. \tag{B.1}$$

Let $(W_g, Y_{B+A_{\min}}(\cdot), \ldots, Y_{B+A_{\max}}(\cdot), E(\cdot), X)$ be a draw from $\mathbb{P}^{\mathcal{D}}_{|g}$. For each $a \in [A_{\min}, A_{\max}]$ define $Y_{B+a} := Y_{B+a}(W_g)$, $E := E(W_g)$. The observed data are $(Y_{B+A_{\min}}, \ldots, Y_{B+A_{\max}}, E, X, W_g)$.

This assumption implies that the policy variable W_g is independent of unit-level potential outcomes and fixed attributes but has a g-specific distribution. The identification in this setup reduces to accounting for the differences in the distribution of W_g across groups. Below, we will discuss several scenarios of how this can be done. Note that the resulting probability model is g-specific, and we do not put any probability measure on the set $\mathcal{G}$ itself. As a result, the probability model can be viewed as a fixed population one, with groups playing the role of the population. Alternatively, it can be viewed as a superpopulation model where we condition on the realized groups.

In what follows, for brevity we use $\mathbb{E}_g[\cdot]$ to denote the g-specific expectation conditional on the realized W_g, e.g,

$$\mathbb{E}_g[Y_{B+a}|E, X] := \mathbb{E}[Y_{B+a}|E, X, W_g].$$

For each group g the conditional expectation $\mathbb{E}_g[Y_{B+a}|E, X]$ is a g-specific random variable.

Remark B.1. Assumption B.1 implicitly fixes all other g-specific factors related to W_g. Such conditioning, in principle, can make the distribution of $\mathbb{P}^W_{|g}$ extremely complicated. To see this, suppose that $W_g = (W_{g,1}, \ldots, W_{g,T_{\max}})$ and there is another, unobserved shock $U_g = (U_{g,1}, \ldots, U_{g,T_{\max}})$, which is also relevant for the potential outcomes. For simplicity suppose that each $(W_{g,t}, U_{g,t}) \sim \mathcal{N}(0, \Sigma)$, independently across t. The marginal distribution of $W_{g,t}$ is simple, $W_{g,t} \sim \mathcal{N}(0, \sigma^2_w)$, and given data from multiple groups or a long time series, we can learn σ^2_w. However, the conditional distribution is very complicated: $W_{g,t} \,|\, (U_{g,t} = u_{g,t}) \sim \mathcal{N}(\beta u_{g,t}, \sigma^2_{cond})$. In particular, for each g and t, the mean is different and equal to $\beta u_{g,t}$. We cannot hope to learn these parameters of the conditional

distribution regardless of how many groups and periods we observed. This problem does not arise in randomized experiments, in which the experimental protocol specifies the distribution of W_g and guarantees that W_g is independent of any unobservables.

B.2 Measurement model and event times

Using the notation introduced previously, we state our next assumption.

Assumption B.2. (Measurement model)
For any $g \in \mathcal{G}$, any $t \in [B + A_{\min}, B + A_{\max}]$ and any w we have:

$$Y_t(w) = \alpha(w) + \lambda_{g,t}(X, w) + \sum_{h \geq 0} \tau_h^e(w)\{t - E(w) = h\}\{E(w) = e\} + \varepsilon_t(w),$$
$$\mathbb{E}[\varepsilon_t(w)|\alpha(w), \boldsymbol{\tau}(w), X, E(w)] = 0,$$

where $\alpha(w)$ and $\{\tau_h^e(w)\}_{e,h}$, $\varepsilon_t(w)$ are random variables with g-specific distributions, and $\lambda_{g,t}(\cdot)$ is a deterministic function.

This assumption imposes a particular structure on the potential outcomes. We can interpret it as policy invariance: The relationship between the objects we want to measure, $\tau_h^e(w)$, and the outcomes does not depend on the realized policy.

Our next assumption restricts the relationship between $E(w)$ and $\tau_h^e(w)$.

Assumption B.3. (Event times)
Either (a) for each $g \in \mathcal{G}$ and any w we have $E(w) \equiv E$, or (b) for each $g \in \mathcal{G}$ and any w we have

$$E(w) \perp\!\!\!\perp \tau_h^e(w)|X.$$

This assumption guarantees that the policy of interest either does not influence the event times or the event times are conditionally random within the subpopulations defined by X. The first assumption is appropriate in cases in which there is no observed reaction of E to policy. Conversely, the second assumption is pertinent if we consider $E(w)$ as a random event rather than a decision influenced by other parameters of the model. Specifically, if $E(w)$ is determined in an experiment, the last statement holds by design.

B.3 Identification

We fix two user-specified parameters, $h_0 \leq 0$ and $h_1 \geq 0$— the minimum and maximum horizon over which we aim to analyze ULES, respectively. We make the following assumption.

Assumption B.4. (Full support)
Random variable $E - B$ has full conditional support, i.e., for any $g \in \mathcal{G}$, w, and $a \in \{A_{\min}, \ldots, A_{\max}\}$, $\mathbb{E}[\mathbf{1}\{E(w) - B = a\}|X] > 0$ X-a.s.

This simplifying full-support assumption does not affect the identification logic but makes the exposition and statistical analysis more straightforward. In particular, it allows us to define a relevant set of relative event times, $A(h_0, h_1) := \{A_{\min} - h_0 + 1, \ldots, A_{\max} - h_1 + h_0 - 1\}$ independently of X.

For a given g, $h_0 \leq h \leq h_1$ and $a \in A(h_0, h_1)$ we have

$$\left(\lambda_{g,B+a+h}(X) - \frac{1}{a - A_{\min} + h_0} \sum_{l=A_{\min}}^{a-1+h_0} \lambda_{g,B+l}(X)\right) =$$
$$\mathbb{E}\left[\left(Y_{E+h} - \frac{1}{E - B - A_{\min} + h_0} \sum_{l=A_{\min}}^{E-1+h_0} Y_{B+l}(X)\right) \frac{\mathbf{1}\{E - B > a + h_1 - h_0\}}{\mathbf{E}[\mathbf{1}\{E - B > a + h_1 - h_0\}|X]}|X\right],$$

where the RHS is well-defined thanks to Assumption B.4. Finally, for a given g and h, we define

$$\hat{\tau}_h := Y_{E+h} - \frac{1}{E - B - A_{\min} + h_0} \sum_{l=A_{\min}}^{E-B-1+h_0} Y_{B+l} -$$
$$\left(\lambda_{g,E+h}(X) - \frac{1}{E - B - A_{\min} + h_0} \sum_{l=A_{\min}}^{E-1+h_0} \lambda_{g,B+l}(X)\right) = \tau_h^E(W_g) + \nu_h^E,$$

as long as $E - B \in A(h_0, h_1)$. The following lemma describes the properties of $\hat{\tau}_h$.

Lemma 1. *Suppose Assumptions B.1 - B.4 hold. Then for any $g \in \mathcal{G}$ and $0 \leq h \leq h_1$, and $a \in A(h_0, h_1)$ we have*

$$\mathbb{E}_g\left[\frac{\hat{\tau}_h \mathbf{1}\{E - B = a\}}{\mathbb{E}[\mathbf{1}\{E - B = a\}|X]}|X\right] = \begin{cases} \mathbb{E}_g[\tau_h^{B+a}(W_g)|E - B = a, X], & \text{if part (a) of Assumption B.3 holds,} \\ \mathbb{E}_g[\tau_h^{B+a}(W_g)|X], & \text{if part (b) of Assumption B.3 holds.} \end{cases}$$

Proof. Using Assumption B.1,B.2 and B.4 and the definition of $\hat{\tau}_h$ we have

$$\mathbb{E}_g\left[\frac{\hat{\tau}_h \mathbf{1}\{E - B = a\}}{\mathbb{E}[\mathbf{1}\{E - B = a\}|X]}|E, X\right] = \mathbb{E}_g\left[\frac{(\tau_h^E(W_g) + \nu_h)\mathbf{1}\{E - B = a\}}{\mathbb{E}[\mathbf{1}\{E - B = a\}|X]}|E, X\right] =$$
$$\mathbb{E}_g\left[\frac{\tau_h^{B+a}(W_g)\mathbf{1}\{E - B = a\}}{\mathbb{E}[\mathbf{1}\{E - B = a\}|X]}|E, X\right].$$

If part (a) of Assumption B.3 holds we have by definition

$$\mathbb{E}_g\left[\frac{\tau_h^{B+a}(W_g)\mathbf{1}\{E - B = a\}}{\mathbb{E}[\mathbf{1}\{E - B = a\}|X]}|X\right] = \mathbb{E}_g\left[\tau_h^{B+a}(W_g)|E - B = a, X\right].$$

Alternatively, if part (b) of Assumption B.3 holds, then we have

$$\mathbb{E}_g\left[\frac{\tau_h^{B+a}(W_g)\mathbf{1}\{E - B = a\}}{\mathbb{E}[\mathbf{1}\{E - B = a\}|X]}|X\right] = \mathbb{E}_g\left[\frac{\tau_h^{B+a}(W_g)\mathbf{1}\{E(W_g) - B = a\}}{\mathbb{E}[\mathbf{1}\{E(W_g) - B = a\}|X]}|X\right] =$$
$$\mathbb{E}_g\left[\tau_h^{B+a}(W_g)|X\right]$$

□

This result, while straightforward, shows the different implications of part of Assumption B.3. For instance, the first part does not allow us to distinguish between state dependence and selection, while the second does.

B.3.1 Known $\mathbb{P}^W_{|g}$

In this subsection, we focus on environments in which the $\mathbb{P}^W_{|g}$ distribution is either known by design or can be identified from the available data. This assumption is natural if W_g is assigned in an experiment (perhaps with a g-specific design), and thus its distribution is known to the analyst. Alternatively, suppose we assume that $\mathbb{P}^W_{|g}$ does not vary across g, and we have access to data from many groups. In that case, we can identify the parameters of the distribution of W_g by pooling the information across groups. The latter structure is analogous to standard analysis under unconfoundedness (see Imbens and Rubin (2015) for the textbook treatment), where the distribution of the treatment variable is unknown but is constant across well-defined subpopulations.

To state our next proposition, we introduce additional notation and define the objects of interest. Let $\gamma(w)$ be a bounded function of w such that $\int \gamma(w)dw = 0$ and $\int w\gamma(w)dw = 1$, where the integrals are computed using a relevant measure on the domain of W (e.g., counting in case W is discrete or Lebesgue if it is continuous) and define

$$\delta^e_h(\gamma) := \int \gamma(w)\tau^e_h(w)dw.$$

To understand the logic behind this quantity, suppose $w \in \{0, 1\}$. Then the only possible contrasts are $\gamma(1) = -\gamma(0) = 1$ and $\delta^e_h(\gamma)$ is the conventional unit-specific treatment effect $\tau^e_h(1) - \tau^e_h(0)$. Expectations of the objects of this type will be our estimands of interest. In particular, we consider two types of expectations:

$$\delta^e_h(\gamma, x, e) := \mathbb{E}[\delta^e_h(\gamma)|X = x, E = e]$$

and

$$\delta^e_h(\gamma, x) := \mathbb{E}[\delta^e_h(\gamma)|X = x].$$

Observe that in $\delta^e_h(\gamma, x, e)$, dependence on e appears twice: as the index of the relevant event time and the index of the subpopulation, we compute the expectation for.

Proposition 1. (Identification with known $\mathbb{P}^W_{|g}$)
Suppose Assumptions B.1 - B.4 hold. Suppose that $\mathbb{P}^W_{|g}$ is known and has a density $f_g(\cdot)$ (with respect to the relevant measure on the domain of W). Then for any contrast $\gamma(\cdot)$ we have for each group $g \in \mathcal{G}$, $0 \le h \le h_1$, and $a \in A(h_0, h_1)$

$$\mathbb{E}\left[\frac{\gamma(W_g)}{f_g(W_g)}\frac{\hat{\tau}_h \mathbf{1}\{E - B = a\}}{\mathbb{E}[\mathbf{1}\{E - B = a\}|X]}\right] = \begin{cases} \mathbb{E}\left[\delta^{B+a}_h(\gamma, X, B + a)\right], & \text{if part (a) of Assumption B.3 holds,} \\ \mathbb{E}[\delta^{B+a}_h(\gamma, X)], & \text{if part (b) of Assumption B.3 holds.} \end{cases}$$

as long as $|\gamma(w)\tau^e_h(w)| \le Z$ and $\mathbb{E}[Z] < \infty$.

Proof. Using the results of Lemma 1 we have if part (a) of Assumption B.3 holds

$$\begin{aligned} &\mathbb{E}\left[\frac{\gamma(W_g)}{f_g(W_g)}\frac{\hat{\tau}_h \mathbf{1}\{E - B = a\}}{\mathbb{E}[\mathbf{1}\{E - B = a\}|X]}\right] = \\ &\mathbb{E}\left[\mathbb{E}\left[\frac{\gamma(W_g)}{f_g(W_g)}\tau^{B+a}_h(W_g)|E - B = a, X\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\delta^{B+a}_h(\gamma, X, B + a)|E - B = a, X\right]\right] = \\ &\mathbb{E}\left[\delta^{B+a}_h(\gamma, X, B + a)\right], \end{aligned}$$

where we used Fubini's theorem to get the second equality. Alternatively, if part (b) of Assumption B.3 holds,

then we get, again using Lemma 1 and Assumption B.1,

$$\mathbb{E}\left[\frac{\gamma(W_g)}{f_g(W_g)}\frac{\hat{\tau}_h \mathbf{1}\{E-B=a\}}{\mathbb{E}[\mathbf{1}\{E-B=a\}|X]}\right] = \mathbb{E}\left[\frac{\gamma(W_g)}{f_g(W_g)}\mathbb{E}_g[\tau_h^{B+a}(W_g)|X]\right] = \mathbb{E}[\delta_h^{B+a}(\gamma, X)].$$

□

Proposition 1 provides a group-specific moment that identifies a relevant causal effect. In the statistical analysis, we can aggregate these moment conditions (across groups) to construct an unbiased estimator for an average (across groups) causal estimand.

B.3.2 Partially known $\mathbb{P}_{|g}^W$

This section considers environments in which $\mathbb{P}_{|g}^W$ is not known, nor can it be identified from the data. This premise is reasonable in observational studies, in which the g-specific probability distribution of W_g can never be fully learned unless we make restrictive assumptions. Instead, we will impose restrictions on certain features of $\mathbb{P}_{|g}^W$.

In our analysis, we will focus on a practically relevant case in which W_g has a time-series structure:

$$W_g = (W_{g,1}, \ldots, W_{g,T_{\max}}).$$

Given this structure, we restrict the causal model for $\tau_h^e(w)$.

Assumption B.5. (Linear time-homogeneous dynamic model)
For any e, h and w we have for each $g \in \mathcal{G}$

$$\tau_{g,h}^e(w) = \beta_h^e + \sum_{j=0}^{h+1} \delta_{j,h} w_{e+h-j},$$

where β_h^e and $\{\delta_{j,h}\}_{j=0}^{h+1}$ are g-specific random variables.

This model imposes several restrictions on the underlying potential outcomes. The first is linearity: We assume that the interactions of policies from different periods do not matter for potential outcomes. Second, we impose a dynamic structure, assuming only policy levels from periods $e-1, \ldots, e+h$ are relevant. This imposes the non-anticipation assumption (future policies are irrelevant to current outcomes) and limited dynamics (lags in distant periods are irrelevant). The final restriction is time homogeneity: The effect of the policy does not structurally change with e—i.e., there is no $\delta_{j,h}^e$.

We make the following assumption about the structure of $\mathbb{P}_{|g}^W$.

Assumption B.6. (Restricted mean)
For any $g \in \mathcal{G}$ and t we have

$$\mathbb{E}[W_{g,t}] = a_g + b_t.$$

This assumption imposes a particular model for the mean of $W_{g,t}$. Note that if we observe a finite number of periods or a finite number of groups, then the parameters $\{a_g, b_t\}_{g,t}$ cannot be identified from the data. As a result, we cannot learn even the first moment of $\mathbb{P}_{|g}^W$ from the data.

For each h and e we define $\boldsymbol{\delta}_h^\top := (\delta_{0,h}, \ldots, \delta_{h+1,h})$ and

$$\left(\Delta \boldsymbol{W}_{g,h}^e\right)^\top := (W_{g,e+h} - W_{g,e+h-1}, \ldots, W_{g,e-1} - W_{g,e-2}).$$

Also, for each $h < h_1$, $g \in \mathcal{G}$ and e we define

$$\Delta\hat{\tau}_h^{B+a} := \hat{\tau}_h \left(\frac{\mathbf{1}\{E - B = a\}}{\mathbb{E}[\mathbf{1}\{E - B = a\}|X]} - \frac{\mathbf{1}\{E - B = a - 1\}}{\mathbb{E}[\mathbf{1}\{E - B = a - 1\}|X]} \right)$$

The next proposition provides us with a g-specific moment condition.

Proposition 2. (IDENTIFICATION WITH PARTIALLY KNOWN $\mathbb{P}_{|g}^W$)
Suppose Assumptions B.1, B.2, and B.4 - B.6 hold. Fix $0 \leq h \leq h_1$. Suppose either part (a) of Assumption B.3 holds and $\mathbb{E}[\boldsymbol{\delta}_h|X, E] = \mathbb{E}[\boldsymbol{\delta}_h|X]$ or, alternatively, part (b) of Assumption B.3 holds. Then for any $\{a, a-1\} \in A(h_0, h_1)$ we have

$$\mathbb{V}[\Delta \boldsymbol{W}_{g,h}^{B+a}|X]\boldsymbol{\delta}_{g,h}(X) = \mathbb{E}[(\Delta \boldsymbol{W}_{g,h}^{B+a} - \mathbb{E}[\Delta \boldsymbol{W}_{g,h}^{B+a}])\Delta\hat{\tau}_h^{B+a}|X],$$

where $\boldsymbol{\delta}_{g,h}(X) := \mathbb{E}[\boldsymbol{\delta}_h|X = x]$.

Proof. Using Lemma 1 we get if part (a) of Assumption B.3 is satisfied:

$$\mathbb{E}_g\left[\Delta\hat{\tau}_h^{B+a}|X\right] = \mathbb{E}_g\left[\tau_h^{B+a}(W_g)|X, E - B = a\right] - \mathbb{E}_g\left[\tau_h^{B+a-1}(W_g)|X, E - B = a - 1\right].$$

Next, using Assumption B.5 and the premise of the proposition, we get

$$\begin{aligned}
&\mathbb{E}_g\left[\tau_h^{B+a}(W_g)|X, E - B = a\right] - \mathbb{E}_g\left[\tau_h^{B+a-1}(W_g)|X, E - B = a - 1\right] = \\
&\mathbb{E}_g\left[\beta_h^{B+a}|X, E - B = a\right] - \mathbb{E}_g\left[\beta_h^{B+a-1}|X, E - B = a - 1\right] + \delta_{g,h}^{\top}(X)\Delta \boldsymbol{W}_{g,h}^{B+a}.
\end{aligned}$$

The rest of the proof follows using Assumption B.1 and B.6. Alternatively, if part (b) of Assumption B.3 is satisfied, the computation is more straightforward, and we have using Lemma 1

$$\mathbb{E}_g\left[\Delta\hat{\tau}_h^{B+a}|X\right] = \mathbb{E}_g\left[\tau_h^{B+a}(W_g) - \tau_h^{B+a-1}(W_g)|X\right].$$

The rest then follows in the same way as before. □

This result shows that we can use the knowledge of the first two moments of $\Delta W_{g,t}$ to construct a moment for the average value of the relevant coefficients. If part (b) of Assumption B.3 is satisfied, then this result does not rely on additional restrictions. However, if part (b) of Assumption B.3 is satisfied, then to achieve the same result we need to impose restrictions on the underlying heterogeneity and assume $\mathbb{E}[\boldsymbol{\delta}_h|X, E] = \mathbb{E}[\boldsymbol{\delta}_h|X]$.

To understand why Proposition 2 is useful in practice, suppose that we observe independent data from multiple groups $g \in \mathcal{G}$.[23] Then we have for any $\{a, a-1\} \in A(h_0, h_1)$

$$\begin{aligned}
&\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}(\Delta \boldsymbol{W}_{g,h}^{b+a} - \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\Delta \boldsymbol{W}_{g,h}^{b+a})\Delta\tau_{g,h}^{b+a}(x, W_g) = \\
&\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\mathbb{E}[(\Delta \boldsymbol{W}_{g,h}^{b+a} - \mathbb{E}[\Delta \boldsymbol{W}_{g,h}^{b+a}])\Delta\tau_{g,h}^{b+a}(x, W_g)] + O_p\left(\frac{1}{\sqrt{|\mathcal{G}|}}\right) = \\
&\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\mathbb{V}[\Delta \boldsymbol{W}_{g,h}^{b+a}]\boldsymbol{\delta}_{g,h} + O_p\left(\frac{1}{\sqrt{|\mathcal{G}|}}\right),
\end{aligned}$$

[23] In the discussion below, we rely on the statistical model described in the next section that postulates that policy variables are independent across groups.

where $\Delta\tau_{g,h}^{b+a}(x, W_g) := \mathbb{E}_g[\Delta\hat{\tau}_h^{B+a}|X = x, E - B = a]$. Similarly, we have

$$\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}(\Delta\boldsymbol{W}_{g,h}^{b+a} - \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\Delta\boldsymbol{W}_{g,h}^{b+a})(\Delta\boldsymbol{W}_{g,h}^{b+a} - \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\Delta\boldsymbol{W}_{g,h}^{b+a})^\top = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\mathbb{V}[\Delta\boldsymbol{W}_{g,h}^{b+a}] + O_p\left(\frac{1}{\sqrt{|\mathcal{G}|}}\right).$$

As a result:

$$\begin{aligned}\hat{\boldsymbol{\delta}}_h(x) := \left(\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}(\Delta\boldsymbol{W}_{g,h}^{b+a} - \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\Delta\boldsymbol{W}_{g,h}^{b+a})(\Delta\boldsymbol{W}_{g,h}^{b+a} - \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\Delta\boldsymbol{W}_{g,h}^{b+a})^\top\right)^{-1} \times \\ \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}(\Delta\boldsymbol{W}_{g,h}^{b+a} - \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\Delta\boldsymbol{W}_{g,h}^{b+a})\Delta\tau_{g,h}^{b+a}(x, W_g) = \\ \left(\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\mathbb{V}[\Delta\boldsymbol{W}_{g,h}^{b+a}]\right)^{-1} \times \left(\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\mathbb{V}[\Delta\boldsymbol{W}_{g,h}^{b+a}]\boldsymbol{\delta}_{g,h}(x)\right) + O_p\left(\frac{1}{\sqrt{|\mathcal{G}|}}\right). \quad \text{(B.2)}\end{aligned}$$

This result is directly useful if $\boldsymbol{\delta}_{g,h}(x) = \boldsymbol{\delta}_h(x)$—i.e., there is no heterogeneity in averages across groups. If such heterogeneity exists, we need to strengthen Assumption B.6 and either assume that $\mathbb{V}[\Delta\boldsymbol{W}_{g,h}^{b+a}]$ does not vary over g or that it can be estimated. In the latter case, we can learn the average effect (across groups) by using $\mathbb{V}[\Delta\boldsymbol{W}_{g,h}^{b+a}]^{-1}(\Delta\boldsymbol{W}_{g,h}^{b+a} - \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\Delta\boldsymbol{W}_{g,h}^{b+a})$ as instruments; see Goldsmith-Pinkham et al. (2024) for further details and alternative procedures.

Even if $\mathbb{V}[\Delta\boldsymbol{W}_{g,h}^{b+a}]$ is known or can be identified, Proposition 2 is not directly useful if part (a) of Assumption B.3 holds but $\mathbb{E}[\boldsymbol{\delta}_h|X, E] \neq \mathbb{E}[\boldsymbol{\delta}_h|X]$—i.e., there is systematic heterogeneity in coefficients across units with different observed event times. To address this case, we need to use moments in levels analogously to Arellano and Bover (1995):

$$\mathbb{E}[(\Delta\boldsymbol{W}_{g,h}^{b+a} - \mathbb{E}[\Delta\boldsymbol{W}_{g,h}^{b+a}])\tau_{g,h}^{b+a}(x)] = \mathbb{E}[(\Delta\boldsymbol{W}_{g,h}^{b+a} - \mathbb{E}[\Delta\boldsymbol{W}_{g,h}^{b+a}])(\boldsymbol{W}_{g,h}^{b+a})^\top]\boldsymbol{\delta}_{g,h}(x, b+a).$$

Again, if either (a) $\boldsymbol{\delta}_{g,h}(x, e)$ does not vary over g or (b) $\mathbb{E}[(\Delta\boldsymbol{W}_{g,h}^{b+a} - \mathbb{E}[\Delta\boldsymbol{W}_{g,h}^{b+a}])(\boldsymbol{W}_{g,h}^{b+a})^\top]$ can be identified from the data, this moment restriction can be directly used to construct an estimator.

There are two problems with this approach in practice. First, to use it we need to assume that moments $\mathbb{E}[(\Delta\boldsymbol{W}_{g,h}^{b+a} - \mathbb{E}[\Delta\boldsymbol{W}_{g,h}^{b+a}])(\boldsymbol{W}_{g,h}^{b+a})^\top]$ are well behaved and the corresponding matrices can be inverted. Second, this IV approach does not have the double-robustness property discussed by Arkhangelsky and Imbens (2022); Arkhangelsky et al. (2024a), unlike other moments discussed above. See the discussion in Section B.5.

Remark B.2. Assumption B.6 specified a particular model for $\mathbb{E}[W_{g,t}]$, but in applications other models can be more appropriate. For instance, we can consider a dynamic model of the type analyzed by Holtz-Eakin et al. (1988) and write

$$W_{g,t} = b_t + \alpha_g\psi_t + \rho W_{g,t-1} + \nu_{g,t}, \quad \mathbb{E}[\nu_{g,t}(1, W_{g,t-1}, \dots)] = 0,$$

where α_g, b_t, ψ_t, and ρ are unknown parameters. This model implies that for each g the mean satisfies the following recursive formula:

$$\mathbb{E}[W_{g,t}] = b_t + \alpha_g\psi_t + \rho\mathbb{E}[W_{g,t-1}],$$

with unrestricted initial conditions. It is thus substantially more general than the model in Assumption B.6. However, this additional generality does not present a major problem for identification. Consider the following

transformation:

$$\tilde{W}_{g,t} := W_{g,t} - \rho W_{g,t-1} - \frac{\psi_t}{\psi_{t-1}}(W_{g,t-1} - \rho W_{g,t-2}) = b_t - \frac{\psi_t}{\psi_{t-1}} b_{t-1} + \nu_{g,t} - \frac{\psi_t}{\psi_{t-1}} \nu_{g,t-1}.$$

Holtz-Eakin et al. (1988) demonstrate that ρ and $\frac{\psi_t}{\psi_{t-1}}$ are identified from the data on multiple groups and thus $\tilde{W}_{g,t}$ can be constructed. Applying the same transformation to our causal model, we can show a significant generalization of Proposition 2.

Remark B.3. Assumption B.6 is directly connected to the literature on formula instruments (Borusyak and Hull, 2023, 2024; Borusyak et al., 2024a). The assumption that the first moment of the appropriate transformation of $W_{g,t}$ is known (identified) is needed to recenter the instrument. The restrictions for the second moment are required to address the contamination bias, as defined by Goldsmith-Pinkham et al. (2024). The above discussion connects these ideas with the identification results for panel data models.

B.4 Statistical analysis

B.4.1 Statistical model

The statistical model is based on the probability model described above. We assume that the group-specific shocks are realized independently for each group, and then n_g units (with n_g being deterministic) are randomly sampled from each group. Formally, the distribution of the realized data takes the following form:

$$\mathbb{P}^{\mathcal{D}} = \times_{g \in \mathcal{G}} \left(\mathbb{P}^{W}_{|g} \times \left(\times_{i=1}^{n_g} \mathbb{P}^{D(\cdot)}_{|g} \right) \right).$$

The total number of observed units is thus $n := \sum_{g \in \mathcal{G}} n_g$. For a given unit i, we let $g(i)$ be the group from which the data for this unit were generated.

Finally, for each unit i, we convert the g-specific random variables defined before into i-specific random variables through the mapping $i \to g(i)$. For instance, $\mathbb{E}_{g(i)}[Y_{i,B_i+a}|E_i, X_i]$ is an i-specific random variable equal to $\mathbb{E}[Y_{B+a}|E = E_i, X = X_i, W_g = W_{g(i)}]$, where $g = g(i)$ and (Y_{B+a}, E, X, W_g) is an independent random draw from $\mathbb{P}^{\mathcal{D}}_{|g}$.

B.4.2 Estimation

The first step of our analysis relies on constructing $\hat{\tau}_{i,h}$. We do this in two steps following the approach described by Borusyak et al. (2024b). First, we estimate

$$(\hat{\alpha}_i, \hat{\lambda}_{g,t}(\cdot)) := \arg \min_{(\alpha_i, \lambda_{g,t}(\cdot))} \sum_{(i,t): E_i \geq t - h_0} (Y_{i,t} - \alpha_i - \lambda_{g(i),t}(X_i))^2,$$

Since our covariates are discrete, we consider a completely unrestricted class of (g,t)-specific functions $\lambda_{g,t}(\cdot)$. Next, we compute for $h \in \{h_0, \dots, h_1\}$:

$$\hat{\tau}^{BJS}_{i,h} := Y_{i,E_i+h} - \hat{\alpha}_i - \hat{\lambda}_{g(i),t}(X_i).$$

We use $\mathcal{S}_{tr}$ to denote the sample of units for which we can construct $\hat{\tau}^{BJS}_{i,h}$ for all values of $h \in \{h_0, \dots, h_1\}$ and use $\hat{\boldsymbol{\tau}}^{BJS}_i$ to denote the $(h_1 - h_0 + 1)$-dimensional ULES.

B.4.3 Statistical results

Unweighted regression: In the analysis in this section, we maintain part (a) of Assumption B.3. Define $\boldsymbol{W}^e_{g,h} := (W_{g,e+h}, \ldots, W_{g,e-1})$. For each $h \in \{0, \ldots, h_1\}$ our estimator takes the following form:

$$\left(\hat{\boldsymbol{\delta}}_h, \hat{\alpha}_h(\cdot), \hat{\beta}_h(\cdot)\right) := \arg \min_{\alpha(\cdot),\beta(\cdot),\boldsymbol{\delta}_h} \sum_{i \in \mathcal{S}_{tr}} \left(\hat{\tau}^{BJS}_{i,h} - \alpha(E_i, X_i) - \beta(g(i), X_i) - \boldsymbol{\delta}_h^\top \boldsymbol{W}^e_{g,h}\right)^2. \tag{B.3}$$

Define

$$\hat{\pi}_g := \frac{n_g}{n}, \quad \hat{\pi}_{x|g} := \frac{\sum_i \mathbf{1}\{X_i = x, g(i) = g\}}{n_g}, \quad \hat{\pi}_{a|x,g} := \frac{\sum_{i \in \mathcal{S}_{tr}} \mathbf{1}\{X_i = x, g(i) = g, E_i - B_i = a\}}{\sum_i \mathbf{1}\{X_i = x, g(i) = g\}},$$

with the convention that $\hat{\pi}_{a|x,g} = 0$ if $\sum_i \mathbf{1}\{X_i = x, g(i) = g\} = 0$. Next, we define

$$\hat{\tau}^{b+a}_{g,h}(x, W_g) := \frac{\sum_{i:\in \mathcal{S}_{tr}, g(i)=g, X_i=x, E_i - B_i = a} \hat{\tau}^{BJS}_{i,h}}{\sum_{i \in \mathcal{S}_{tr}} \mathbf{1}\{X_i = x, g(i) = g, E_i - B_i = a\}}$$

Formally this object is well-defined only if $\sum_{i \in \mathcal{S}_{tr}} \mathbf{1}\{X_i = x, g(i) = g, E_i - B_i = a\} > 0$, but we can ignore this because it is multiplied by $\hat{\pi}_{a|x,g}$.

It is straightforward to see that the problem (B.3) is equivalent to the following one:

$$\left(\hat{\boldsymbol{\delta}}_h, \hat{\alpha}_h(\cdot), \hat{\beta}_h(\cdot)\right) :=$$
$$\arg \min_{\alpha(\cdot),\beta(\cdot),\boldsymbol{\delta}_h} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|x,g} \left(\hat{\tau}^{b+a}_{g,h}(x, W_g) - \alpha(b+a, x) - \beta(g, x) - \boldsymbol{\delta}_h^\top \boldsymbol{W}^{b+a}_{g,h}\right)^2.$$

For any g, x and $a \in A(h_0, h_1)$ define the conditional probability:

$$\hat{\pi}^c_{a|g,x} := \frac{\hat{\pi}_{a|g,x}}{\sum_{l \in A(h_0,h_1)} \hat{\pi}_{l|g,x}},$$

with the convention that $\hat{\pi}^c_{a|g,x} = 0$ if $\sum_{l \in A(h_0,h_1)} \hat{\pi}_{l|g,x} = 0$.

We use this probability to define the following projection:

$$\tilde{\boldsymbol{W}}^{b+a}_{g,h}(x) := \boldsymbol{W}^{b+a}_{g,h} - \sum_{l \in A(h_0,h_1)} \hat{\pi}^c_{l|g,x} \boldsymbol{W}^{b+l}_{g,h} - \mathbb{E}\left[\boldsymbol{W}^{b+a}_{g,h} - \sum_{l \in A(h_0,h_1)} \hat{\pi}^c_{l|g,x} \boldsymbol{W}^{b+l}_{g,h} | \{X_i, E_i\}_{i \in g}\right].$$

We also consider the following prediction problem:

$$\hat{\boldsymbol{\alpha}}_h(\cdot) := \arg \min_{\boldsymbol{\alpha}_h(\cdot)} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|x,g} \left\| \tilde{\boldsymbol{W}}^{b+a}_{g,h}(x) - \boldsymbol{\alpha}_h(b+a, x) + \sum_{l \in A(h_0,h_1)} \hat{\pi}^c_{l|g,x} \boldsymbol{\alpha}_h(b+l, x) \right\|_2^2,$$

and define the corresponding residual:

$$\hat{\boldsymbol{W}}^{b+a}_{g,h}(x) = \tilde{\boldsymbol{W}}^{b+a}_{g,h}(x) - \left(\hat{\boldsymbol{\alpha}}_h(b+a, x) - \sum_{l \in A(h_0,h_1)} \hat{\pi}^c_{l|g,x} \hat{\boldsymbol{\alpha}}_h(b+l, x)\right).$$

For each (g, x, a) we define

$$\tau_{g,h}^{b+a}(x, W_g) := \mathbb{E}_g[\tau_h^{B+a} | X = x, E - B = a], \quad \beta_{g,h}^{b+a}(x) := \mathbb{E}_g[\beta_h^{B+a} | X = x, E - B = a].$$

Define the error:

$$\zeta_{g,h}^{b+a}(x) := \hat{\tau}_{g,h}^{b+a}(x, W_g) - \tau_{g,h}^{b+a}(x, W_g).$$

For each (g, x, a) we define $\tilde{\zeta}_{g,h}^{b+a}(x)$ and $\hat{\beta}_{g,h}^{b+a}(x)$ analogously to $\tilde{\boldsymbol{W}}_{g,h}^{b+a}(x)$ and $\hat{\boldsymbol{W}}_{g,h}^{b+a}(x)$ above.

We define the following set of functions

$$\mathcal{F} := \Big\{ \boldsymbol{\nu}_{g,h}(\cdot) : \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0, h_1)} \hat{\pi}_{a|g,x} \|\boldsymbol{\nu}_{g,h}(b + a, x)\|_2^2 = 1,$$
$$\nu_{g,h}(b + a, x) = \boldsymbol{\alpha}_h(b + a, x) - \sum_{l \in A(h_0, h_1)} \hat{\pi}_{a|g,x}^c \boldsymbol{\alpha}_h(b + l, x), \text{ for some } \boldsymbol{\alpha}_h(b + l, x) \Big\}$$

Heuristically, for each h these are functions that have a unit norm in the L^2 space generated by the empirical measure on (g, x, a) and have a special structure generated by the underlying functions $\boldsymbol{\alpha}_h(b + a, x)$. Next, we define the conditional variance

$$\Sigma_{g,h}^{b+a}(x) := \mathbb{E}\left[\tilde{\boldsymbol{W}}_{g,h}^{b+a}(x) \left(\tilde{\boldsymbol{W}}_{g,h}^{b+a}(x) \right)^\top | \{E_i, X_i\}_{i \in g} \right],$$

which we use to define another variance, which will be the variance of the OLS estimator,

$$\Sigma := \left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0, h+1)} \hat{\pi}_{a|g,x} \Sigma_{g,h}^{b+a}(x) \right)^{-1} \times$$
$$\left(\sum_g \hat{\pi}_g^2 \mathbb{V}\left[\sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0, h_1)} \hat{\pi}_{a|g,x} \tilde{\boldsymbol{W}}_{g,h}^{b+a}(x) (\hat{\beta}_{g,h}^{b+a}(x) + \hat{\zeta}_{g,h}^{b+a}(x)) | \{E_i, X_i\}_{i \in g} \right] \right) \times$$
$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0, h+1)} \hat{\pi}_{a|g,x} \Sigma_{g,h}^{b+a}(x) \right)^{-1}.$$

We make the following high-level assumption:

Assumption B.7. (HIGH-LEVEL RESTRICTIONS)
As n and $|\mathcal{G}|$ approach infinity the following statements hold conditionally on $\{E_i, X_i\}_{i=1}^n$ with probability approaching 1: (a) the fixed effects do not overfit

$$\sup_{\boldsymbol{\nu}_{g,h}^e(x) \in \mathcal{F}} \left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0, h_1)} \hat{\pi}_{a|g,x} \left(\boldsymbol{\nu}_{g,h}(b + a, x) \right)^\top \tilde{\boldsymbol{W}}_{g,h}^{b+a} \right) = o_p(1);$$

(b) the conditional LLN holds

$$\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0, h_1)} \hat{\pi}_{a|g,x} \tilde{\boldsymbol{W}}_{g,h}^{b+a}(x) \left(\tilde{\boldsymbol{W}}_{g,h}^{b+a}(x) \right)^\top = \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0, h_1)} \hat{\pi}_{a|g,x} \Sigma_{g,h}^{b+a}(x) + o_p(1);$$

(c) the effective sample size is going to infinity

$$\|\Sigma\|_{op} = o_p(1);$$

and (d) the conditional CLT holds

$$\left(\sum_g \hat{\pi}_g^2 \mathbb{V}\left[\sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \tilde{\boldsymbol{W}}_{g,h}^{b+a}(x)(\hat{\beta}_{g,h}^{b+a}(x) + \hat{\zeta}_{g,h}^{b+a}(x))|\{E_i, X_i\}_{i=1}^n\right]\right)^{-\frac{1}{2}} \times$$
$$\sum_g \hat{\pi}_g \hat{\sum}_x \pi_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \tilde{\boldsymbol{W}}_{g,h}^{b+a}(x)(\hat{\beta}_{g,h}^{b+a}(x) + \hat{\zeta}_{g,h}^{b+a}(x)) \to_d \mathcal{N}(0, \mathcal{I}).$$

This assumption is very mild, and we expect it to hold as $|\mathcal{G}|$ approaches infinity even if the group size stays fixed as long as some mild regularity conditions are satisfied. The first part of this assumption says that two-way fixed effects do not asymptotically overfit the relevant policy variation. It is trivially satisfied if the support of X and $E - B$ does not increase with n and $|\mathcal{G}|$. The second part is the requirement that the conditional law of large numbers applies. A sufficient condition for this is that $\sum_g \hat{\pi}_g^2 = o(1)$ and boundedness of $W_{g,t}$. The third restriction is a very mild requirement for the effective sample size, with $\sum_g \pi_g^2 = o(1)$ being a sufficient condition as long as $W_{g,t}$ and potential outcomes are bounded. Finally, the last restriction is the conditional validity of the central limit theorem, which we expect to hold under mild conditional moment restrictions.

Theorem 1. (Asymptotic behavior)
Suppose Assumptions B.1 - B.2, B.4 - B.7 hold and part (a) of Assumption B.3 holds. Suppose for all g we have $\mathbb{E}[\boldsymbol{\delta}_h|X, E] = \boldsymbol{\delta}$. Then, as n and $|\mathcal{G}|$ go to infinity we have

$$\hat{\Sigma}^{-\frac{1}{2}}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \to_d \mathcal{N}(0, \mathcal{I}),$$

where

$$\hat{\Sigma} := \left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x) \left(\hat{\boldsymbol{W}}_{g,h}^{b+a}(x)\right)^\top\right)^{-1} \times$$
$$\left(\sum_g \hat{\pi}_g^2 \left(\sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|x,g} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x)(\hat{\tau}_{g,h}^e(x, W_g) - \hat{\boldsymbol{\delta}}_h \hat{\boldsymbol{W}}_{g,h}^{b+a}(x))\right)^\top \times\right.$$
$$\left.\left(\sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x)(\hat{\tau}_{g,h}^{b+a}(x, W_g) - \hat{\boldsymbol{\delta}}_h \hat{\boldsymbol{W}}_{g,h}^{b+a}(x))\right)\right) \times$$
$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x) \left(\hat{\boldsymbol{W}}_{g,h}^{b+a}(x)\right)^\top\right)^{-1}.$$

Proof. In the proof, we use the fact that $W_{g,t}$ is independent of $\{E_i, X_i\}_{i=1}^n$, and thus, we can treat the latter as fixed. We then conduct the analysis for the subset of realizations of $\{E_i, X_i\}_{i=1}^n$ on which Assumption B.7 holds, and show that for each such realization, we have convergence in distribution. Since this set has a probability approaching one, it implies that the same convergence holds unconditionally.

Using the standard FWL representation of the regression coefficients, we get the following expression for $\hat{\boldsymbol{\delta}}_h$:

$$\hat{\boldsymbol{\delta}}_h =$$
$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x) \left(\hat{\boldsymbol{W}}_{g,h}^{b+a}(x)\right)^\top\right)^{-1} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x) \hat{\tau}_{g,h}^{b+a}(x, W_g) =$$
$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x) \left(\hat{\boldsymbol{W}}_{g,h}^{b+a}(x)\right)^\top\right)^{-1} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x) \tau_{g,h}^{b+a}(x, W_g) +$$
$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x) \left(\hat{\boldsymbol{W}}_{g,h}^{b+a}(x)\right)^\top\right)^{-1} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x) \zeta_{g,h}^{b+a}(x).$$

Using the second part of Assumption B.7, we have for the denominator

$$\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x) \left(\hat{\boldsymbol{W}}_{g,h}^{b+a}(x)\right)^\top =$$
$$\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \tilde{\boldsymbol{W}}_{g,h}^{b+a}(x) \left(\tilde{\boldsymbol{W}}_{g,h}^{b+a}(x)\right)^\top -$$
$$\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \times$$
$$\left(\hat{\boldsymbol{\alpha}}_h(b+a,x) - \sum_{l\in A(h_0,h_1)} \hat{\pi}_{a|g,x}^c \hat{\boldsymbol{\alpha}}_h(b+l,x)\right)\left(\hat{\boldsymbol{\alpha}}_h(b+a,x) - \sum_{l\in A(h_0,h_1)} \hat{\pi}_{a|g,x}^c \hat{\boldsymbol{\alpha}}_h(b+l,x)\right)^\top =$$
$$\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \Sigma_{g,h}^{b+a}(x) + o_p(1) -$$
$$\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \times$$
$$\left(\hat{\boldsymbol{\alpha}}_h(b+a,x) - \sum_{l\in A(h_0,h_1)} \hat{\pi}_{a|g,x}^c \hat{\boldsymbol{\alpha}}_h(b+l,x)\right)\left(\hat{\boldsymbol{\alpha}}_h(b+a,x) - \sum_{l\in A(h_0,h_1)} \hat{\pi}_{a|g,x}^c \hat{\boldsymbol{\alpha}}_h(b+l,x)\right)^\top.$$

We can bound the remaining error

$$\Big\|\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \times$$
$$\left(\hat{\boldsymbol{\alpha}}_h(b+a,x) - \sum_{l\in A(h_0,h_1)} \hat{\pi}_{a|g,x}^c \hat{\boldsymbol{\alpha}}_h(b+l,x)\right)\left(\hat{\boldsymbol{\alpha}}_h(b+a,x) - \sum_{l\in A(h_0,h_1)} \hat{\pi}_{a|g,x}^c \hat{\boldsymbol{\alpha}}_h(b+l,x)\right)^\top \Big\|_{op} \leq$$
$$\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a\in A(h_0,h_1)} \hat{\pi}_{a|g,x} \|\hat{\boldsymbol{\alpha}}_h(b+a,x) - \sum_{l\in A(h_0,h_1)} \hat{\pi}_{a|g,x}^c \hat{\boldsymbol{\alpha}}_h(b+l,x)\|_2^2,$$

and by using standard OLS logic, we have

$$\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \|\hat{\boldsymbol{\alpha}}_h(b+a,x) - \sum_{l \in A(h_0,h_1)} \hat{\pi}^c_{a|g,x} \hat{\boldsymbol{\alpha}}_h(b+l,x)\|_2^2 \leq$$
$$4 \sup_{\boldsymbol{\nu}_{g,h}(x) \in \mathcal{F}} \left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \left(\boldsymbol{\nu}_{g,h}(x)\right)^\top \boldsymbol{W}^e_{g,h} \right) = o_p(1),$$

where the last equality is guaranteed by the first part of Assumption B.7.

The rest follows trivially. First, we have

$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}^{b+a}_{g,h}(x) \left(\hat{\boldsymbol{W}}^{b+a}_{g,h}(x)\right)^\top \right)^{-1} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}^{b+a}_{g,h}(x) \zeta^{b+a}_{g,h}(x) =$$
$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \Sigma^{b+a}_{g,h}(x) + o_p(1) \right)^{-1} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \tilde{\boldsymbol{W}}^{b+a}_{g,h}(x) \hat{\zeta}^{b+a}_{g,h}(x).$$

We can decompose the other part of the estimator:

$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}^{b+a}_{g,h}(x) \left(\hat{\boldsymbol{W}}^{b+a}_{g,h}(x)\right)^\top \right)^{-1} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}^{b+a}_{g,h}(x) \tau^{b+a}_{g,h}(x, W_g) =$$
$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}^{b+a}_{g,h}(x) \left(\hat{\boldsymbol{W}}^{b+a}_{g,h}(x)\right)^\top \right)^{-1} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}^{b+a}_{g,h}(x) \left(\boldsymbol{W}^{b+a}_{g,h}\right)^\top \boldsymbol{\delta}_{g,h}(x) +$$
$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}^{b+a}_{g,h}(x) \left(\hat{\boldsymbol{W}}^{b+a}_{g,h}(x)\right)^\top \right)^{-1} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}^{b+a}_{g,h}(x) \beta^{b+a}_{g,h}(x)$$

For the target, we use that the parameter is not heterogeneous, $\mathbb{E}[\boldsymbol{\delta}_h | X, E] = \boldsymbol{\delta}$, for all g:

$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}^{b+a}_{g,h}(x) \left(\hat{\boldsymbol{W}}^{b+a}_{g,h}(x)\right)^\top \right)^{-1} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}^{b+a}_{g,h}(x) \left(\boldsymbol{W}^{b+a}_{g,h}\right)^\top \boldsymbol{\delta}_{g,h}(x) =$$
$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}^{b+a}_{g,h}(x) \left(\hat{\boldsymbol{W}}^{b+a}_{g,h}(x)\right)^\top \right)^{-1} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}^{b+a}_{g,h}(x) \left(\boldsymbol{W}^{b+a}_{g,h}\right)^\top \boldsymbol{\delta}_h =$$
$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}^{b+a}_{g,h}(x) \left(\hat{\boldsymbol{W}}^{b+a}_{g,h}(x)\right)^\top \right)^{-1} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}^{b+a}_{g,h}(x) \left(\hat{\boldsymbol{W}}^{b+a}_{g,h}(x)\right)^\top \boldsymbol{\delta}_h =$$
$$\boldsymbol{\delta}_h$$

For the final part, we have

$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x) \left(\hat{\boldsymbol{W}}_{g,h}^{b+a}(x)\right)^\top\right)^{-1} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x) \beta_{g,h}^{b+a}(x) =$$
$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \Sigma_{g,h}^{b+a}(x) + o_p(1)\right)^{-1} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \tilde{\boldsymbol{W}}_{g,h}^{b+a}(x) \hat{\beta}_{g,h}^{b+a}(x)$$

We then immediately have, from the third part of Assumption B.7,

$$\Sigma^{-\frac{1}{2}}(\hat{\boldsymbol{\delta}}_h - \boldsymbol{\delta}_h) \to_d \mathcal{N}(0, \mathcal{I}).$$

Since $\|\Sigma\|_{op} \to o_p(1)$, we also have that $\hat{\delta} - \delta = o_p(1)$. It then follows using standard arguments that

$$\hat{\Sigma}^{-1} \times \Sigma = \mathcal{I} + o_p(1).$$

This concludes the proof. □

Theorem 1 abstracts away from all heterogeneity in effects. The limiting distribution has an additional term if such heterogeneity is present. We can compute the probability limit of the estimator with such heterogeneity:

$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x) \left(\hat{\boldsymbol{W}}_{g,h}^{b+a}(x)\right)^\top\right)^{-1} \times$$
$$\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \hat{\boldsymbol{W}}_{g,h}^{b+a}(x) \left(\boldsymbol{W}_{g,h}^{b+a}\right)^\top \boldsymbol{\delta}_{g,h}(x) =$$
$$\left(\sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \Sigma_{g,h}^{b+a}(x)\right)^{-1} \sum_g \hat{\pi}_g \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} \Sigma_{g,h}^{b+a}(x) \boldsymbol{\delta}_{g,h}(x) + o_p(1)$$

Observe that even if $\boldsymbol{\delta}_{g,h}(x) = \boldsymbol{\delta}_h(x)$ for all $g \in \mathcal{G}$ and $\mathbb{V}[\boldsymbol{W}_{g,h}^{b+a}]$ does not depend on g, the resulting estimand is not equal to a convex combination of $\boldsymbol{\delta}_h(x)$, unlike in our analysis in the previous section. The reason for such behavior is that the regression (B.3) induces particular (g, x, a)-specific weights that render the analysis more complicated. To address this problem, we can use alternative weights analogous to the estimator we discussed in Section 2.4. In particular, consider the following weighted regression:

$$\left(\hat{\boldsymbol{\delta}}_h^{WOLS}, \hat{\alpha}_h^{WOLS}(\cdot), \hat{\beta}_h^{WOLS}(\cdot)\right) := \arg \min_{\alpha(\cdot),\beta(\cdot),\boldsymbol{\delta}_h} \sum_{i \in \mathcal{S}_{tr}} \left(\hat{\tau}_{i,h}^{BJS} - \alpha(E_i, X_i) - \beta(g(i), X_i) - \boldsymbol{\delta}_h^\top \boldsymbol{W}_{g,h}^e\right)^2 \frac{1}{\hat{\pi}_{(E_i - B_i)|g(i),X_i}},$$

which is analogous to one we propose in situations where part (b) of Assumption B.3 holds. A straightforward extension of the previous analysis shows that the limit is now equal to

$$\left(\sum_g \hat{\pi}_g \Sigma_{g,h}\right)^{-1} \sum_g \hat{\pi}_g \Sigma_{g,h} \left(\sum_x \hat{\pi}_{x|g} \boldsymbol{\delta}_{g,h}(x)\right)$$

where

$$\Sigma_{g,h} := \sum_x \hat{\pi}_{x|g} \sum_{a \in A(h_0,h_1)} \mathbb{V}\left[\boldsymbol{W}_{g,h}^{b+a} - \frac{1}{|A(h_0,h_1)|} \sum_{l \in A(h_0,h_1)} \boldsymbol{W}_{g,h}^{l} \right].$$

In particular, now if either (a) $\sum_x \hat{\pi}_{x|g} \boldsymbol{\delta}_{g,h}(x) = \boldsymbol{\delta}_h$, or (b) $\Sigma_{g,h} = \Sigma_h$, we have that the resulting estimand is either (a) equal to $\boldsymbol{\delta}_h$, or (b) equal to

$$\left(\sum \hat{\pi}_g\right)^{-1} \sum \hat{\pi}_g \left(\sum_x \hat{\pi}_{x|g}(x) \boldsymbol{\delta}_{g,h}(x) \right).$$

Weighted regression: We now briefly discuss the differences in the analysis for the environments in which part (b) of Assumption B.3 holds. The weighted version of the OLS problem has the following form

$$\left(\hat{\boldsymbol{\delta}}_h^{WOLS}, \hat{\alpha}_h^{WOLS}(\cdot), \hat{\beta}_h^{WOLS}(\cdot)\right) := \arg \min_{\alpha(\cdot),\beta(\cdot),\boldsymbol{\delta}_h} \sum_{i \in \mathcal{S}_{tr}} \left(\hat{\tau}_{i,h}^{BJS} - \alpha(E_i, X_i) - \beta(g(i), X_i) - \boldsymbol{\delta}_h^\top \boldsymbol{W}_{g,h}^{e} \right)^2 \frac{1}{\hat{\pi}_{(E_i - B_i)|g(i),X_i}}$$

This construction implicitly assumes that $\hat{\pi}_{(E_i-B_i)|g(i),X_i} > 0$ for all i—a random event. We define the corresponding indicator variables

$$A := \mathbf{1}\left\{ \min_{g \in \mathcal{G}, x, a \in A(h_0,h_1)} \hat{\pi}_{a|g,x} > 0 \right\}.$$

Following the same path as in the previous case, we wish to conduct the analysis fixing $\{X_i\}_{i=1}^n$ and $A = 1$. However, because E_i is causally related to $W_{g,t}$ the second condition affects the distribution of $\{W_{g,t}\}_t$, shifting its first two moments, which were relevant for our computations. However, as long as $1 - \mathbb{E}[A|\{X_i\}_i] = o_p\left(\sum_{g \in \mathcal{G}} \hat{\pi}_g^2\right)$, we can rely on the following CS bound:

$$|\mathbb{E}[f(\{X_i\}_{i \in g}, \{W_{g,t}\}_t | A = 1, \{X_i\}_i] - \mathbb{E}[f(\{X_i\}_{i \in g}, \{W_{g,t}\}_t)|\{X_i\}_i]| \leq$$
$$\frac{\sqrt{\mathbb{V}[f(\{X_i\}_{i \in g}, \{W_{g,t}\}_t | \{X_i\}_{i=1}^n] \mathbb{E}[A|\{X_i\}_i](1 - \mathbb{E}[A|\{X_i\}_i])}}{\mathbb{E}[A|\{X_i\}_i]} =$$
$$o_p\left(\sqrt{\mathbb{V}[f(\{X_i\}_{i \in g}, \{W_{g,t}\}_t | \{X_i\}_{i=1}^n] \sum_g \hat{\pi}_g^2} \right).$$

It follows that under mild technical conditions (e.g., boundedness of all relevant random variables), fixing $A = 1$ does not change the relevant moments of $W_{g,t}$ enough to affect the first-order analysis. In particular, one can proceed analogously to the previous case, establishing an analog of Theorem 1.

B.4.4 Normalization

Our empirical analysis sometimes normalizes the estimated $\hat{\boldsymbol{\tau}}_i^{BJS}$ by the average imputed outcome. In particular, for each (h, e, x, g) we construct

$$\hat{Y}_{h,e,x,g} := \frac{\sum_{i \in \mathcal{S}_{tr}: E_i = e, X_i = x, g(i) = g} (\hat{\alpha}_i + \hat{\lambda}_{g(i),t}(X_i))}{n_g^{tr}(e,x)},$$

where $n_g^{tr}(e,x) := \sum_{i \in \mathcal{S}_{tr}} \mathbf{1}\{E_i = e, X_i = x, g(i) = g\}$, and then define for each $i \in \mathcal{S}_{tr}$

$$\tilde{\tau}_{i,h}^{BJS} := \frac{\hat{\tau}_{i,h}^{BJS}}{\hat{Y}_{h,E_i,X_i,g(i)}}.$$

Remark B.4. Our theoretical and statistical analysis is conducted for $\hat{\tau}^{BJS}_{i,h}$, ignoring the normalization. To address the normalization, one needs to restate Assumption B.5 for the normalized version of $\tau^e_h(\omega)$. Statistical guarantees analogous to those we establish below would hold for the normalized coefficients as long as the number of units per group is sufficiently large. In particular, group-level clustering would still deliver correct inference.

B.5 Alternative analysis

The analysis in this section relied on random variation in W_g to establish causal claims. However, a large body of applied work, particularly in policy evaluation, focuses on conditional analysis that fixes the path of W_g and instead relies on unobserved shocks. For instance, this is the case for the standard DiD analysis. This subsection briefly discusses how we can incorporate such analysis into our framework.

If we fix W_g, our previous probability model does not allow for group-level uncertainty. To incorporate such uncertainty, we assume that in addition to W_g—the observed policy shock—the potential outcomes are also affected by the unobserved policy shocks U_g. We extend Assumption B.1 and now assume

$$(E, X, Y_{B+A_{\min}}(\cdot), \ldots, Y_{B+A_{\max}}(\cdot)) \perp\!\!\!\perp (W_g, U_g),$$

where, as before, each variable can have a g-specific distribution. The measurement model remains the same as in Assumption B.2, but now incorporates dependence on (u):

$$Y_t(w,u) = \alpha(w,u) + \lambda_{g,t}(X,w,u) + \sum_{h\geq 0} \tau^e_h(w,u)\{E(w,u) - t = h\}\{E(w,u) = e\} + \varepsilon_t(w,u),$$
$$\mathbb{E}[\varepsilon_t(w,u)|\alpha(w,u), \boldsymbol{\tau}(w,u), X, E] = 0.$$

The next restriction extends part (a) of Assumption B.3:

$$E(w,u) \equiv E.$$

We can also consider a straightforward extension of part (b) of Assumption B.3; it does not affect the discussion below much, and we focus on a simpler case for brevity.

These restrictions guarantee

$$\tau^e_{g,h}(x,w,u) := \mathbb{E}[\tau^e_h(w,u)|X = x, E = e, W_g = w, U_g = u] = \mathbb{E}[\tau^e_h(w,u)|X = x, E = e].$$

We assume that U_g has the same structure as W_g, and thus we can write

$$U_g = (U_{g,1}, \ldots, U_{g,T_{\max}}).$$

We use this structure to extend Assumption B.5:

$$\tau^e_h(w,u) = \beta^e_h + \sum_{j=0}^{h+1} \beta_{1,j,h} u_{e+h-j} + \sum_{j=0}^{h+1} \delta_{j,h} w_{e+h-j},$$

where β^e_h, $\beta_{1,j,h}$, and $\delta_{j,h}$ are g-specific random variables. So far, this model is a generalization of the one we had

previously, but next, we impose restrictions that are more demanding than before. In particular, we assume

$$\begin{aligned}
&\mathbb{E}[\beta_h^e | X = x, E = e] = \beta_{0,1,g,h}(x) + \beta_{0,2,h}(x, e), \\
&\mathbb{E}[\beta_{1,j,h} | X = x, E = e] = \beta_{1,j,h}(x), \\
&\mathbb{E}[\delta_{j,h} | X = x, E = x] = \delta_{g,j,h}(x).
\end{aligned}$$

Here, the last restriction is the same as we imposed before, and the first two are new. Finally, we restrict the conditional distribution of $U_{g,t}$ and assume

$$U_{g,t} = a_g + b_t + \nu_{g,t}, \quad \mathbb{E}[\nu_{g,t} | W_g] = 0.$$

Using all these restrictions, we arrive at the following model:

$$\tau_{g,h}^e(x, W_g, U_g) = \tilde{\beta}_{1,g,h}(x) + \tilde{\beta}_{2,h}(x, e) + \sum_{j=0}^{h+1} \delta_{g,j,h}(x) W_{g,e+h-j} + \tilde{\nu}_{g,e+h}, \quad \mathbb{E}[\tilde{\nu}_{g,e+h} | W_g] = 0, \tag{B.4}$$

where

$$\begin{aligned}
\tilde{\beta}_{1,g,h}(x) &:= \beta_{0,1,g,h}(x) + \sum_{j=0}^{h+1} \beta_{1,j,h}(x) a_g, \\
\tilde{\beta}_{2,h}(x, e) &= \beta_{0,2,h}(x, e) + \sum_{j=0}^{h+1} \beta_{1,j,h}(x) b_{e+h-j}, \\
\tilde{\nu}_{g,e+h} &:= \sum_{j=0}^{h+1} \beta_{1,j,h}(x) \nu_{g,e+h-j}.
\end{aligned}$$

Compared with the model we had before, this one has a specific structure of the error term, which was not previously present. It is directly related to the one considered by De Chaisemartin and d'Haultfoeuille (2020).

To understand identification and estimation in this model, we consider a simple example with three periods, which is the same as in the main text but now allows for covariates. Suppose that W_g takes three possible values, $W_g \in \{(0,0,0), (0,0,1), (0,1,1)\}$. Consider the following difference:

$$\Delta\tau_{g,0}^2(x, W_g, U_g) = \tau_{g,0}^2(x, W_g, U_g) - \tau_{g,0}^1(x, W_g, U_g) = \Delta\tilde{\beta}_{2,h}(x) + \sum_{j=0}^{1} \delta_{g,j,0}(x) \Delta W_{g,2-j} + \Delta\tilde{\nu}_{g,2}.$$

Using the restriction on $\Delta\tilde{\nu}_{g,2}$ we get the following:

$$\begin{aligned}
\hat{\delta}_0(x) := &\frac{\sum_{g \in \mathcal{G}} \Delta\tau_{g,0}^2(x, W_g, U_g) \{\Delta W_{g,2} = 1, \Delta W_{g,1} = 0\}}{\sum_{g \in \mathcal{G}} \{\Delta W_{g,2} = 1, \Delta W_{g,1} = 0\}} - \\
&\frac{\sum_{g \in \mathcal{G}} \Delta\tau_{g,0}^2(x, W_g, U_g) \{\Delta W_{g,2} = 0, \Delta W_{g,1} = 0\}}{\sum_{g \in \mathcal{G}} \{\Delta W_{g,2} = 0, \Delta W_{g,1} = 0\}} = \\
&\frac{\sum_{g \in \mathcal{G}} \delta_{g,0,0}(x) \{\Delta W_{g,2} = 1, \Delta W_{g,1} = 0\}}{\sum_{g \in \mathcal{G}} \{\Delta W_{g,2} = 1, \Delta W_{g,1} = 0\}} + O_p\left(\frac{1}{\sqrt{|\mathcal{G}|}}\right),
\end{aligned}$$

and similarly

$$\hat{\delta}_1(x) := \frac{\sum_{g\in\mathcal{G}} \Delta\tau^2_{g,0}(x, W_g, U_g)\{\Delta W_{g,2} = 0, \Delta W_{g,1} = 1\}}{\sum_{g\in\mathcal{G}}\{\Delta W_{g,2} = 0, \Delta W_{g,1} = 1\}} - \frac{\sum_{g\in\mathcal{G}} \Delta\tau^2_{g,0}(x, W_g, U_g)\{\Delta W_{g,2} = 0, \Delta W_{g,1} = 0\}}{\sum_{g\in\mathcal{G}}\{\Delta W_{g,2} = 0, \Delta W_{g,1} = 0\}} = \frac{\sum_{g\in\mathcal{G}} \delta_{g,1,0}(x)\{\Delta W_{g,2} = 0, \Delta W_{g,1} = 1\}}{\sum_{g\in\mathcal{G}}\{\Delta W_{g,2} = 0, \Delta W_{g,1} = 1\}} + O_p\left(\frac{1}{|\mathcal{G}|}\right).$$

As a result, we can construct an estimator for a particular weighted average of the underlying coefficients. Note that the weights are random, and their expectation is not generally identified. It is also easy to see that this estimator is equivalent to estimating (B.4) by OLS with two-way fixed effects using data on $\mathcal{G}$ groups and is also equivalent to the estimator we considered previously in (B.2). This result is directly connected to the ideas developed by Arkhangelsky et al. (2024a).

Remark B.5. The discussion in this section illustrates the benefits and costs of the conditional approach. On the positive side, we do not need to restrict the marginal distribution of W_g, i.e., we have not used an analog of Assumption B.6. However, instead, we assumed the presence of unobserved shocks with a two-way structure and mean-independent errors, as well as more substantial restrictions on heterogeneity. Finally, in terms of the estimand, we got a convex combination of average effects, but the weights are different for different coefficients, which makes the interpretation more challenging.

B.6 Validation

Our empirical analysis relies on the first part of Assumption B.3. To validate this assumption, we conduct empirical analysis based on the model described in this section. For each t, g, w, b define the following conditional probability (hazard function):

$$\pi_{g,t}(w, b) := \mathbb{E}[E(w) = t | E(w) > t - 1, B = b].$$

We assume that this probability has the following functional form:

$$\pi_{g,t}(w, b) = \pi^0_{g,t}(b) + \beta_g(t - b)w_{t-1}.$$

This specification does not restrict the baseline variation in the hazard rate without policy. Still, it assumes that the potential probability only depends on the current (at $t - 1$) level of policy, and this effect is constant across time, though it varies across age. Define the following two variables:

$$\Delta\pi_{g,t}(w, b) := \pi_{g,t}(w, b) - \pi_{g,t-1}(w, b - 1),$$
$$\Delta\pi^0_{g,t}(b) := \pi^0_{g,t}(b) - \pi^0_{g,t-1}(b - 1).$$

Using this definition, we arrive at the following equation:

$$\Delta\pi_{g,t}(W_g, b) = \Delta\pi^0_{g,t}(b) + \beta_g(t - b)\Delta W_{g,t-1}.$$

This implies that under Assumption B.6 we have the following:

$$\frac{\sum_{g\in\mathcal{G}} \Delta\pi_{g,t}(W_g, t-b)\left(\Delta W_{g,t-1} - \frac{1}{\mathcal{G}}\sum_{g\in\mathcal{G}}\Delta W_{g,t-1}\right)}{\sum_{g\in\mathcal{G}}\left(\Delta W_{g,t-1} - \frac{1}{\mathcal{G}}\sum_{g\in\mathcal{G}}\Delta W_{g,t-1}\right)^2} = \sum_{g\in\mathcal{G}}\beta_g(t-b)\omega_{g,t} + O_p\left(\sqrt{\sum_{g\in\mathcal{G}}\omega_{g,t}^2}\right),$$

where $\omega_{g,t} := \frac{\mathbb{V}[\Delta W_{g,t-1}]}{\sum_{g\in\mathcal{G}}\mathbb{V}[\Delta W_{g,t-1}]}$. If none of the group-specific variances dominates, then we have $O_p\left(\sqrt{\sum_{g\in\mathcal{G}}\omega_{g,t}^2}\right) = O_p\left(\frac{1}{\sqrt{|\mathcal{G}|}}\right)$. This implies that the OLS regression of the group-level outcome $\Delta\pi_{g,t}(W_g, \tilde{x}, t-b)$ on the policy differences is a consistent estimator for a weighted average effect of the policy. In practice, we do not observe the probability $\pi_{g,t}(w,b)$ directly; instead, we use its unbiased estimator.

B.7 Examples

We now return to examples from Section 2.5 and discuss them in more detail.

B.7.1 One- vs. two-step estimation

Recall that we consider estimating the following equation

$$Y_{i,t} = \alpha_i + \lambda_{g,t} + \sum_{h\geq 0}(\beta_h + \delta_h W_g)\mathbf{1}\{E_i - t = h\} + \epsilon_{i,t}$$

assuming that W_g is binary and is randomly assigned and $t = 0, 1, 2$. For simplicity, we difference out the unit fixed effects by subtracting the outcomes in the first period, which leads to the following equation

$$\tilde{Y}_{i,t} = \tilde{\lambda}_{g,t} + \sum_{h\geq 0}(\beta_h + \delta_h W_g)\mathbf{1}\{t - E_i = h\} + \tilde{\epsilon}_{i,t},$$

where $\tilde{Y}_{i,t} := Y_{i,t} - Y_{i,0}$ for $t = 1, 2$, and $\tilde{\lambda}_{g,t}$ and $\tilde{\epsilon}_{i,t}$ are defined analogously. We consider the population limit of this problem; that is, we solve for

$$\left(\{\lambda_{g,t}^{OLS}\}_{g,t}, \{\beta_h^{OLS}, \delta_h^{OLS}\}_{h=0}^1\right) = \arg\min_{\{\lambda_{g,t}\}_{g,t}, \{\beta_h,\delta_h\}_{h=0}^1} \sum_{g\in\mathcal{G}}\sum_{t=1}^2 \mathbb{E}\left[\left(\tilde{Y}_t - \lambda_{g,t} - \sum_{h\geq 0}(\beta_h + \delta_h W_g)\mathbf{1}\{t - E = h\}\right)^2\right]$$

First, suppose that part (a) of Assumption (3) holds. Then, by using the FWL theorem multiple times, we get the following:

$$\{\delta_h^{OLS}\}_{h=0}^1 = \arg\min_{\{\delta_h\}_{h=0}^1} \sum_{g\in\mathcal{G}}\sum_{t=1}^2 \mathbb{E}\left[\left(\tau_{g,h}^{t-h}(W_g, E)\mathbf{1}\{t - E = h\} - \sum_{h\geq 0}\delta_h\left(W_g - \frac{1}{2}\right)\left(\mathbf{1}\{t - E = h\} - \pi_{g,t,h}\right)\right)^2\right],$$

where $\pi_{g,t,h} := \mathbb{E}_g[\mathbf{1}\{t - E = h\}]$ and $\tau_{g,h}^{t-h}(W_g, E) := \mathbb{E}_g[\tau_h^{t-h}|E]$. Because W_g is binary we have

$$\tau_{g,h}^{t-h}(W_g, E) = \beta_{g,h}^{t-h}(E) + \delta_{g,h}^{t-h}(E)W_g = \beta_{g,h}^{t-h}(E) + \frac{1}{2}\delta_{g,h}^{t-h}(E) + \delta_{g,h}^{t-h}(E)W_g\left(W_g - \frac{1}{2}\right).$$

Applying the FWL result once again, we get

$$\{\delta_h^{OLS}\}_{h=0}^1 = \arg\min_{\{\delta_h\}_{h=0}^1} \sum_{g\in\mathcal{G}}\sum_{t=1}^{2}\mathbb{E}\left[\left(\sum_{h\geq 0}\left(\delta_{g,h}^{t-h}(E)\mathbf{1}\{t-E=h\} - \delta_h\left(\mathbf{1}\{t-E=h\}-\pi_{g,t,h}\right)\right)\right)^2\right] \Rightarrow$$

$$\begin{pmatrix}\delta_0^{OLS}\\ \delta_1^{OLS}\end{pmatrix} = \left(\sum_{g\in\mathcal{G}}(V_{g,1}+V_{g,2})\right)^{-1}\sum_{g\in\mathcal{G}}\mathbb{E}[V_{g,1}(E)\boldsymbol{\delta}_g^1(E) + V_{g,2}(E)\boldsymbol{\delta}_g^2(E)].$$

where

$$V_{g,t}(E) := \begin{pmatrix}\mathbf{1}\{t-E=0\}(\mathbf{1}\{t-E=0\}-\pi_{g,t,0})\\ \mathbf{1}\{t-E=1\}(\mathbf{1}\{t-E=1\}-\pi_{g,t,1})\end{pmatrix}^\top \begin{pmatrix}\mathbf{1}\{t-E=0\}(\mathbf{1}\{t-E=0\}-\pi_{g,t,0})\\ \mathbf{1}\{t-E=1\}(\mathbf{1}\{t-E=1\}-\pi_{g,t,1})\end{pmatrix}$$

$$\boldsymbol{\delta}_g^t(E) := \begin{pmatrix}\delta_{g,0}^t(E)\\ \delta_{g,1}^{t-1}(E)\end{pmatrix}, \quad V_{g,t} := \mathbb{E}[V_{g,t}(E)]$$

This illustrates that, in general, the OLS coefficients suffer from the contamination bias introduced by Goldsmith-Pinkham et al. (2024). Note that this is the case even if we shut down the heterogeneity in the event times, making $\boldsymbol{\delta}_g^t(E) = \boldsymbol{\delta}_g^t$. This still leaves heterogeneity across groups, which matters even though W_g is i.i.d. across groups.

We now turn to the second part of Assumption 3. To illustrate the point in the most straightforward setting, we focus on a single period; that is, we consider estimating the following linear equation

$$\tilde{Y}_{i,t} = \tilde{\lambda}_g + (\beta_0 + \delta_0 W_g)\mathbf{1}\{E_i = 1\} + \tilde{\epsilon}_{i,t},$$

where $\tilde{Y}_{i,t} := Y_{i,t} - Y_{i,0}$ and $t = 1$.[24] Focusing on the population problem, we have

$$\left(\{\lambda_g^{OLS}\}_g, \beta^{OLS}, \delta^{OLS}\right) = \arg\min_{\{\lambda_g\}_g,\beta,\delta}\sum_{g\in\mathcal{G}}\mathbb{E}\left[\left(\tilde{Y}_t - \lambda_g - (\beta+\delta W_g)\mathbf{1}\{E=1\}\right)^2\right]$$

Applying the FWL theorem, we get that this is equivalent to estimating

$$\left(\beta^{OLS}, \delta^{OLS}\right) = \arg\min_{\beta,\delta}\sum_{g\in\mathcal{G}}\mathbb{E}\left[\left(\tau_{g,0}^0(W_g)(\mathbf{1}\{t-E=h\}-\pi_{g,1,0}) - (\beta+\delta W_g)(\mathbf{1}\{E=1\}-\pi_{g,1,0})\right)^2\right],$$

where we used the fact that $\tau_0^0(w)$ is independent of $E(w)$. The last problem is thus equivalent to the following one

$$\left(\beta^{OLS}, \delta^{OLS}\right) = \arg\min_{\beta,\delta}\sum_{g\in\mathcal{G}}\mathbb{E}\left[\left(\tau_{g,0}^0(W_g) - (\beta+\delta W_g)\right)^2\omega_g(W_g)\right],$$

where $\omega_g(W_g) := \mathbb{V}[\mathbf{1}\{E=1\}|W_g]$. Using the FWL results again, we get that

$$\delta^{OLS} := \frac{\sum_g \mathbb{E}[\omega_g(W_g)\tau_{g,0}^0(W_g)(W_g-\mu)]}{\sum_g \mathbb{E}[\omega_g(W_g)(W_g-\mu)^2]} = \frac{\sum_g \beta_{g,0}^0\mathbb{E}[\omega_g(W_g)(W_g-\mu)]}{\sum_g \mathbb{E}[\omega_g(W_g)(W_g-\mu)^2]} + \frac{\sum_g \delta_{g,0}^0\mathbb{E}[\omega_g(W_g)W_g(W_g-\mu)]}{\sum_g \mathbb{E}[\omega_g(W_g)(W_g-\mu)^2]},$$

where $\mu := \frac{\sum_g \mathbb{E}[\omega_g(W_g)W_g]}{\sum_g \mathbb{E}[\omega_g(W_g)]}$. The g-specific weight $\mathbb{E}[\omega_g(W_g)(W_g-\mu)]$ is in general not equal to zero (these weights sum up to zero across all groups, though), and thus the OLS estimator is biased even if $\delta_{g,0}^0$ is constant across g. This result should not be surprising because the OLS problem that we got uses weights that depend on

[24] With two periods, we will also have the contamination bias in addition to the bias we discuss below.

W_g thus shifting the distribution of W_g away from the uniform.

B.7.2 Unit-level policy variation

Next, we consider the model with unit-level variation in policy:

$$\tilde{Y}_{i,t} = \tilde{\lambda}_t(W_i) + \sum_{h\geq 0}(\beta_h + \delta_h W_i)\mathbf{1}\{t - E_i = h\} + \tilde{\epsilon}_{i,t}.$$

Again, focusing on the limit problem, we have the following optimization problem:

$$\begin{aligned}&(\{\lambda_t^{OLS}(\cdot)\}_{t=1}^2, \{(\beta_h^{OLS}, \delta_h^{OLS})\}_{h=0}^1) = \\ &\arg\min_{\{\lambda_t(\cdot)\}_{t=1}^2, \{(\beta_h,\delta_h)\}_{h=0}^1} \sum_t \mathbb{E}[(Y_t - \tilde{\lambda}_t(W_i) - \sum_{h\geq 0}(\beta_h + \delta_h W)\mathbf{1}\{t - E = h\})^2]\end{aligned}$$

Applying the FWL theorem, we get that this problem is equivalent to the following one:

$$\begin{aligned}&(\{(\beta_h^{OLS}, \delta_h^{OLS})\}_{h=0}^1) = \\ &\arg\min_{\{\{(\beta_h,\delta_h)\}_{h=0}^1} \sum_t \mathbb{E}\left[\left(\sum_{h\geq 0}\tau_h^{t-h}(W,E)\mathbf{1}\{t - E = h\} - \sum_{h\geq 0}(\beta_h + \delta_h W)(\mathbf{1}\{t - E = h\} - \pi_{t,h}(W))\right)^2\right].\end{aligned}$$

where $\pi_{t,h}(W)) := \mathbb{E}[\mathbf{1}\{t - E = h\}|W]$. If part (a) of Assumption (3) holds, then the last problem is equivalent to

$$(\{(\delta_h^{OLS})_{h=0}^1) = \arg\min_{\{(\delta_h)_{h=0}^1} \sum_t \mathbb{E}\left[\left(\sum_{h\geq 0}(\delta_h^{t-h}(E)\mathbf{1}\{t - E = h\} - \delta_h W(\mathbf{1}\{t - E = h\} - \pi_{t,h})\right)^2\right],$$

which, using similar computations as in the previous section, implies

$$\begin{pmatrix}\delta_0^{OLS}\\ \delta_1^{OLS}\end{pmatrix} = ((V_1 + V_2))^{-1}\sum_{g\in\mathcal{G}}\mathbb{E}[V_1(E)\boldsymbol{\delta}^1(E) + V_2(E)\boldsymbol{\delta}^2(E)],$$

where all components are defined analogously to our previous analysis. Even if E is not correlated with the heterogeneity in coefficients, we still get contamination bias because of the variation over time.

Alternatively, if part(b) of Assumption 3 holds, then we get the following problem:

$$(\boldsymbol{\beta}^{OLS}, \boldsymbol{\delta}^{OLS}) = \arg\min_{(\boldsymbol{\beta}^{OLS}, \boldsymbol{\delta}^{OLS})} \sum_t \mathbb{E}[(\boldsymbol{\tau}^t(W) - \boldsymbol{\beta} - \boldsymbol{\delta}W))^\top V_t(W)(\boldsymbol{\tau}^t(W) - \boldsymbol{\beta} - \boldsymbol{\delta}W))],$$

where $(\boldsymbol{\beta}^{OLS})^\top := (\beta_0^{OLS}, \beta_1^{OLS})$, $\boldsymbol{\delta}^{OLS}$ is defined analogously, and $(\boldsymbol{\tau}^t(W))^\top := (\tau_0^1(W), \tau_1^0(W))$, and we define $\boldsymbol{\beta}^t$ and $\boldsymbol{\delta}^t$ analogously. We also defined

$$V_t(W) := \mathbb{V}[\mathbf{1}\{t - E = 0\}, \mathbf{1}\{t - E = 1\}|W]$$

Solving this problem, we get the following:

$$\boldsymbol{\delta}^{OLS} = \left(\sum_{t=1}^{2} V_t(1)\right)^{-1} \left(\sum_t V_t(1)\boldsymbol{\delta}^t\right) + \left(\sum_{t=1}^{2} V_t(1)\right)^{-1} \left(\sum_t V_t(1)\left(\boldsymbol{\beta}^t - \boldsymbol{\beta}^{OLS}\right)\right),$$

where $\boldsymbol{\beta}^{OLS} = \left(\sum_{t=1}^{2} V_t(0)\right)^{-1} \left(\sum_t V_t(0)\boldsymbol{\beta}^t\right)$. As before, there is a selection bias and a contamination bias caused by heterogeneity across periods.

B.7.3 Causal analysis vs. projections

Our computations in the previous section show that if we only use a single period rather than two periods, then the resulting coefficients have a causal interpretation. This section illustrates the same phenomenon using a more straightforward model.

Consider a statistical model where

$$Y_i = \tau_i X_i + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | X_i, W_i] = 0,$$

and researchers want to project τ_i on binary $W_i \in \{0, 1\}$, i.e., estimate the coefficient in the regression

$$\tau_i = \beta + \delta W_i + \nu_i$$

As long as $X_i \neq 0$ almost surely, we can construct $\hat{\tau}_i := \frac{Y_i}{X_i}$ for all units and project it on W_i. This two-step procedure is conceptually analogous to our proposal. Alternatively, we can directly estimate a single linear equation

$$Y_i = \alpha + \beta X_i + \delta X_i W_i + \tilde{\varepsilon}_i,$$

where we substituted τ_i with its projection on W_i. The problem with this approach is that the new error $\tilde{\varepsilon}_i$ can be systematically correlated with X_i, thus potentially rendering the whole analysis invalid. This issue arises regardless of the nature of the variation in W_i, which can be randomly assigned. See Muris and Wacker (2022) for a detailed analysis of a more general version of this problem.

The situation becomes more nuanced if we postulate a causal model for τ_i, and write $\tau_i(W_i)$. In this case, if W_i is randomly assigned, and neither X_i nor ε_i is causally affected by W_i, then the one-step OLS estimator converges to a weighted average effect

$$\mathbb{E}\left[(\tau_i(1) - \tau_i(0))\frac{X_i^2}{\mathbb{E}[X_i^2]}\right].$$

This effect differs from the standard average treatment effect from the two-step regression, but it is still a meaningful causal quantity.

C Household Labor Supply and Childcare Decisions Model

This section presents a simple household model to illustrate the economic intuition behind the relationship between the child penalty and childcare provision in Figure IV, discussed in Section 4.3. Let $U(c, k)$ represent the joint utility function of the household, where the household enjoys consumption c and child-rearing time k. For simplicity, we assume both fathers f and mothers m have the same utility function. The joint household maximization problem can be expressed as:

$$\max_{c,k} U(c, k) = u(c_m, k_m) + u(c_f, k_f)$$

subject to the budget constraint:

$$p \cdot d + c_m + c_f = w_m \cdot h_m + h_f$$

where p is the price of outsourcing child-rearing time d, h_i are market working hours at a wage w_m for mothers and wage for fathers at a wage rate we normalize to be the numeraire $w_f = 1$. Each parent can split their total time of $T = 1$ between market work and child-rearing in the following way:

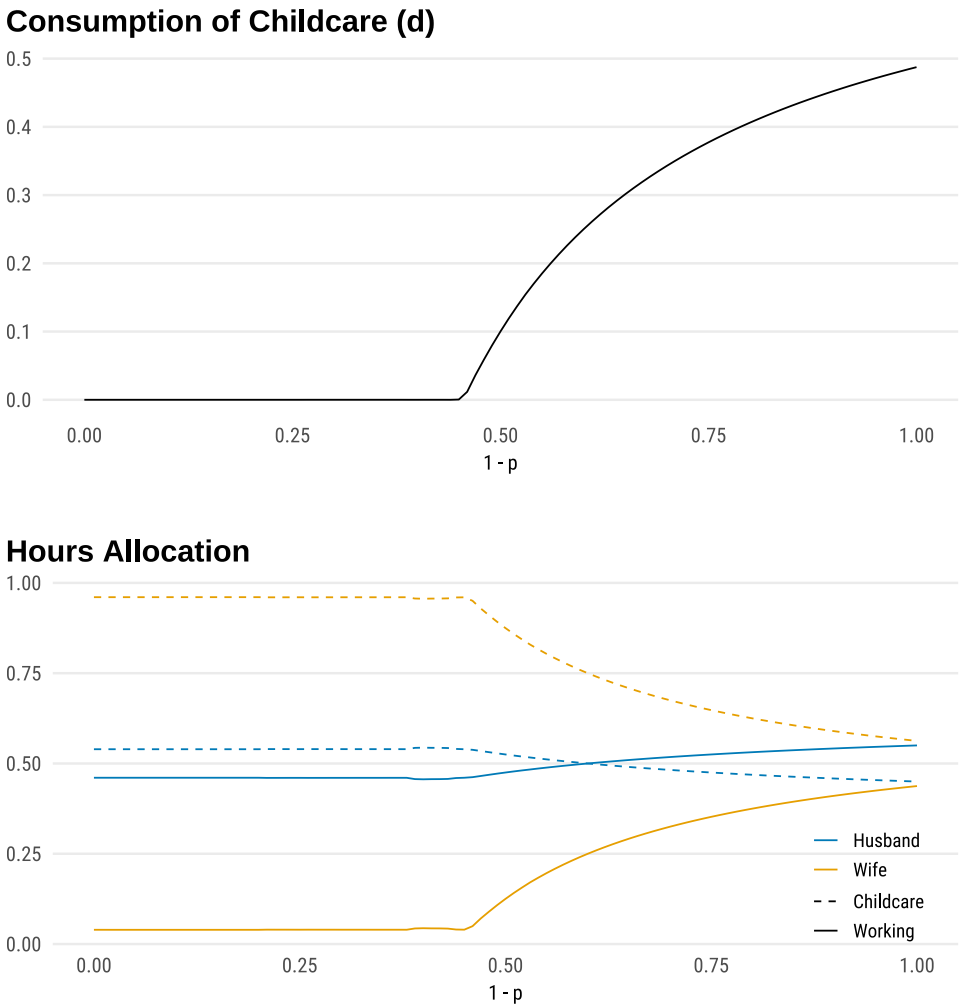
$$1 = h_i + k_i.$$

Finally, someone needs to care for the child daily for a total time of $T = 1$, which can be outsourced by choosing d. Finally, someone needs to do home production for a total time of $\bar{h}$, which results in the following household time constraint:

$$1 + \bar{h} = k_m + k_f + d.$$

This simple model produces a heterogeneous relationship between labor supply and the cost of outsourcing childcare time, as Figure C.1 shows.

Figure C.1: Time allocation to childcare and labor supply versus childcare cost



Notes: This figure presents the XX.