Identification of Child Penalties

Dor Leventer*

November 24, 2025

Job Market Paper

[*Click here for most updated version*]

Abstract

A growing body of research estimates child penalties, the gender gap in the effect of parenthood on earnings, using event studies that normalize treatment effects by counterfactual earnings. I formalize the identification framework underlying this approach and show that it does not identify its target estimand when the parallel trends assumption *in levels* is violated. Insights from human capital theory suggest such violations are likely: higher-ability individuals tend to delay childbirth and have steeper earnings growth, a mechanism that causes conventional estimates to understate child penalties for early-treated parents. I propose instead to target the effect of parenthood on the gender earnings ratio and show that this parameter is identified without additional assumptions. Using Israeli administrative data, a bias-bounding exercise for the conventional estimator suggests substantial understatement (25–50%) of child penalties for early-treated parents, while later-treated groups exhibit smaller biases. I also document that estimates of the new parameter are roughly 30% smaller in absolute value than the conventional estimates for later-treated parents.

The transition to parenthood is central to the onset of gender inequality in labor markets (Goldin, 2024). An important and growing body of research quantifies the gender gap in the effect of parenthood on labor market outcomes using event-study designs that normalize estimated effects post treatment (e.g., Kleven, Landais, & Sogaard, 2019). Recent critiques discuss how event-study based estimates may be biased (Melentyeva & Riedel, 2023), building on a broader econometric literature on biases in two-way fixed-effects regressions with

*Tel Aviv University. I am indebted to my PhD Advisor Itay Saporta-Eksten for guidance in this project. For comments I thank Yoav Goldstien, Analia Schlosser, Roee Levy, Oren Danieli, Ran Abramitzky, David Neumark, Netanel Ben-Porath, and seminar participants in Tel Aviv. I gratefully acknowledge financial support from The Israel Pollak Fellowship Program for Excellence and the Arlozorov Forum Labor Markets Scholarship.

staggered treatment timing (e.g., Goodman-Bacon, 2021). While these critiques focus on biases in estimation, less attention has been paid to the identification framework underlying event studies that rely on post-treatment normalization.

This paper begins by formalizing the identification framework that can be rationalized from normalized event-study empirical practice. I reverse engineer the identification framework from the pre-treatment validation tests of this empirical approach, which I term Normalized Triple Differences (NTD). Surprisingly, I find that this framework does not identify the gender gap in normalized effects, the target causal estimand most studies intend to estimate, when the parallel trends assumption in levels does not hold. To achieve identification, the paper proceeds in two directions. The first examines whether alternative frameworks, namely difference-in-differences (DID) and triple differences (TD), are valid in the child-penalty context. I explore this both theoretically, drawing on insights from human-capital models, and empirically, using pre-treatment validation tests. The second examines whether identification can be achieved under NTD. I show that it can, either by changing the target causal estimand to the effect of parenthood on the gender earnings ratio, or, by adding an assumption on fathers' effects and applying a bias-correction formula. I conclude by discussing aggregation across treatment groups, and specifically how variation in treatment weights shape the interpretation of differences in aggregated estimators. Throughout, I illustrate the theoretical arguments on Israeli administrative data.

I begin by describing the normalized event-study strategy, which proceeds in three steps. First, for each individual, the event is defined as the age at first childbirth, and separate regressions are estimated by gender, regressing the outcome on a set of event time indicators and fixed effects for age and calendar year. Second, predicted outcomes are computed using only the estimated fixed effects, often interpreted as the expected earnings absent treatment. Third, the estimated event-time coefficients are normalized by the mean of the predicted outcomes, expressing effects as proportional changes in earnings relative to the counterfactual without the effect of parenthood. Considering earnings as the outcome of interest, these normalized estimates are interpreted as the effect of parenthood on earnings by gender, and the post-treatment gender gap is typically labeled the child penalty.

To lay the groundwork for the identification results below, I formalize the causal estimands which are targeted in the above normalized event-study approach. The analysis focuses on causal estimands evaluated at a given age and for a single treatment group from a given gender, where treatment is defined by age at first childbirth, and on 2×2 comparisons in which the control group is the nearest not-yet-treated treatment group. Within gender, I consider three causal estimands: the average potential outcome (APO) in the counterfactual without children, the average treatment effect (ATE), and the normalized average

treatment effect, defined as the ratio of the ATE to the counterfactual APO. The normalized average treatment effect serves as the 2×2 building block underlying the above normalized event-time coefficients, which aggregate multiple treatment groups.

I then discuss the first main result of the paper: the identification assumption that is implied by the validation tests used in normalized event studies, and the resulting bias for the gender gap in normalized effects even when the assumption holds. Child-penalty applications commonly show validation tests that in pre-treatment periods the normalized estimates for males and females are similar (see, e.g., Andresen & Nix, 2022; Kleven, Landais, & Sogaard, 2019). Translating this procedure to a 2×2 comparison shows that, if a no-anticipation assumption holds, this type of test validates that violations of parallel trends are equal across genders after normalizing by the counterfactual APOs. I refer to the framework that assumes no anticipation and equal normalized parallel-trend violations across genders as Normalized Triple Differences (NTD). Surprisingly, I find that in post-childbirth periods, the NTD framework does not identify the gender gap in normalized effects unless the parallel-trends assumption in levels also holds, meaning that the trajectories of counterfactual earnings are the same across treatment and control groups. This result implies that the normalized event-study approach relies on an identification framework that, even when its assumptions are satisfied, fails to identify the target causal estimand unless the parallel-trends assumption in levels also holds; an assumption that lies outside the NTD framework and whose validity I examine below.

Having established that the NTD framework does not identify the target causal estimand, the paper proceeds in two directions. The first is to examine whether alternative identification frameworks, DID or TD, can achieve identification in the child penalty context. The second direction is to consider modifications to the NTD framework that restore identification. I begin with the alternative frameworks, and then return to the NTD-based alternatives.

Starting with DID, I analyze the parallel-trends assumption building on insights from canonical behavioral models of fertility and human capital (e.g., Becker et al., 1990; Ben-Porath, 1967). In such models, individuals decide how much to invest in human capital and when to have children based on inherent ability and preferences. These models predict that individuals with higher labor-market ability delay childbirth and invest more in human capital, implying that later treatment groups have steeper earnings trajectories even in the counterfactual without children. Consequently, the counterfactual earnings trajectories of later treatment groups are steeper than those of earlier treatment groups. These differences in earning trends due to selection on treatment suggest a mechanism that causes DID to overstate the earnings loss from parenthood, in 2×2 comparisons where the treatment

group are early-treated and control groups are later treated. This direction of parallel-trends violation also implies that the NTD estimand equals the true gender gap in normalized effects multiplied by a bias term smaller than one; hence, it understates the true effect. Using Israeli administrative data, I document evidence consistent with this type of selection on treatment: parents who have their first child at later ages come from higher-income and more-educated families and perform better on national mathematics exams.[1]

For TD, I discuss violations of its identifying assumption that differences in counterfactual earnings trends are the same for mothers and fathers. Whether this assumption holds is difficult to determine without additional structure, as it depends on how male and female earnings would evolve differently in the absence of childbirth. To address this, I derive a decomposition that expresses the TD violation in terms of the ratio of female to male counterfactual earnings absent children, and show that a sufficient condition for the TD assumption to hold is that this ratio remains constant across treatment groups and ages. Using Israeli earnings data before first childbirth, I document life-cycle patterns of this counterfactual gender earnings ratio and find that it declines with age after 28—indicating widening counterfactual gender inequality—and that later-treated groups are generally more gender-equal at any given age. Combining these regularities with the argument on selection, I show through an empirical exercise that for early-treated parents at later ages, TD is likely to understate the gender gap in the effect of parenthood on earnings.

Concluding the discussion of violations of the identification assumptions, I present validation tests for DID, TD, and NTD. Although NTD does not identify the target causal estimand, I include validation tests for it because below I present alternatives within NTD that achieve identification under this framework. I argue that such tests should be conducted at the 2×2 level: because the control group changes with age, each treatment group has multiple corresponding validation tests—one for each control group used in the post-childbirth periods. This differs from the validation exercises commonly reported in the literature, which aggregate across multiple treatment and control groups. Using the Israeli data, I document two main patterns in pre-treatment ages. First, within gender, DID estimates of pre-treatment differences in trends increase with the gap in treatment timing between the treated and control groups. Second, TD and NTD exhibit larger pre-treatment deviations for later-treated groups, consistent with widening counterfactual gender inequality at older ages.

This concludes the examination of whether alternative frameworks can be used to iden-

[1] To my knowledge, two other studies document similar correlations between age at first childbirth and human capital–related covariates. Melentyeva and Riedel (2023) find a positive correlation with grandparents' education, and Jensen et al. (2024) find a positive correlation with parents' years of schooling. Both patterns are also evident in the Israeli data.

tify child penalties. The next part of the paper discusses two alternatives that under the NTD framework restore identification. First, I define a new causal estimand: the change in the female–male earnings ratio between the observed treatment and the counterfactual without childbirth, which is point identified under NTD without additional assumptions and is directly relevant to the child-penalty question. This result motivates a new estimator, the difference between the gender ratio of mean observed earnings and the gender ratio of DID-based estimated counterfactual APOs, which targets the new causal estimand. For inference, I develop cluster-robust standard errors based on the estimator's influence function. Using the Israeli data, I estimate both the original estimand, the gender gap in normalized effects, and the new estimand, the effect of parenthood on gender inequality. I find that the estimates are similar for early treatment groups, while estimates of the new estimand are smaller in magnitude for later groups. I show that these differences can be attributed to three factors: bias in the original estimand due to violations of parallel trends, the gender ratio in counterfactual earnings, and the magnitude of fathers' effects.

This leads to the second alternative, in which the original NTD causal estimand is, in principle, point identified by imposing an additional identification assumption of a null average treatment effect for fathers, yielding a bias-correction formula. In the above discussion, DID is the only framework that allows identification of effects for specific genders, as opposed to directly targeting gender gaps. As I argue theoretically that parallel trends are violated and show empirically that it fails validation tests, I cannot use DID to estimate the effect for fathers. Hence, I use the bias-correction framework as a diagnostic tool to provide suggestive evidence on the magnitude of bias in NTD estimates, rather than to estimate the causal estimands of interest directly. Applying this approach to the Israeli data suggests that naïve NTD estimators for earlier treatment groups are biased upward toward zero; for example, relative to baseline NTD, allowing fathers' normalized effects to vary from -10% to 10% yields bias-corrected estimates that become 24–51% more negative five years after childbirth for the treatment group aged 26 at first birth.

Finally, I examine aggregation across treatment groups. Considering comparisons of aggregated estimates across strata, such as comparing estimates of different countries or parent types, I argue that variation in the treatment distribution complicates interpretation. For example, OECD data show that the United States has a right-skewed distribution of ages at first birth, while Italy and Spain are more left-skewed, implying greater weight on later-treated groups in U.S. aggregates. I illustrate this empirically by showing that aggregated estimates, constructed from single-treatment-group estimates using the Israeli data, can differ substantially when evaluated under alternative treatment distributions.

This paper is most closely related to the influential literature that estimates the long-

run effect of becoming a parent on gender inequality in the labor market, the so-called child penalty, using administrative data and event-study designs. Bertrand et al. (2010) provide an early and highly influential analysis of earning trajectories among MBAs, documenting the emergence of gender gaps in earnings and labor supply around the first childbirth. Angelov et al. (2016) extend this approach by using population-wide data and longer event horizons in a with-couple event study specification, showing that the effects of parenthood on gender inequality persist for more than a decade after childbirth. Kleven, Landais, and Sogaard (2019) further advance the literature by introducing an approach which normalizes event study coefficients, in order to adjust for gender differences unrelated to childbirth, thereby enabling more meaningful cross-gender comparisons. A growing body of subsequent work builds on this influential framework to explore the mechanisms behind the child penalty (e.g., Andresen & Nix, 2022; Kleven, 2022; Kleven et al., 2021).

I contribute to this literature in two main ways. First, by formalizing the NTD identification framework that normalized event study approach relys on, establishing both the non-identification result for the gender gap in normalized effects and the new identification result for the effect of parenthood on gender inequality. This can be seen as a recommendation for future studies of child penalty and its mechanisms that rely on the normalized event study approach to estimate the new proposed estimator. Second, I discuss how aggregate pre-trend tests, which are the common tests in the literature, are not a sufficient test for the identification assumptions and propose instead to conduct pre-trend diagnostics at the treatment–control pair level. Beyond the child-penalty context, the NTD framework developed here may also be useful in other settings for identifying the effects of a treatment on group inequality, such as the impacts of job loss on gender or racial earnings inequality.

This paper also relates to studies that use contemporary DID estimation methods. As examples, in the child penalty context, Melentyeva and Riedel (2023) implement a stacked DID design (Cengiz et al., 2019; Wing et al., 2024), Fajardo-Gonzalez et al. (2024) and Lin (2025) use the estimator of Sun and Abraham (2021), and Bearth (2024) adopt the approach of Callaway and Sant'Anna (2021). While these methods correct estimation bias stemming from staggered treatment timing, they do not address identification bias due to violations of the parallel trends assumption. I discuss violations of the parallel-trends assumption theoretically, drawing on insights from economic models. Specifically, I argue that differences in human capital likely cause DID to overstate the effect of parenthood on earnings for early treatment groups at later ages. More broadly, the approach of augmenting pre-treatment empirical validation tests with post-treatment theoretical reasoning may be applied in other settings as well.

The rest of the paper is organized as follows. Section 1 presents the normalized event-study

empirical strategy and the data used for illustrating the theory. Section 2 establishes the result that NTD does not identify the gender gap in normalized effects. Section 3 examines whether DID and TD are viable alternative frameworks, and Section 4 presents modifications to NTD that restore identification. Section 5 discusses aggregation across multiple treatment groups, and Section 6 concludes. To facilitate replication and application of the proposed estimators, I developed an open-source R package, `childpen`, which implements the discussed estimators.

1 Empirical Context

The child-penalty literature estimates the gender gap in normalized effects of parenthood on labor-market outcomes using normalized event-study designs and administrative earnings data. This section first formalizes that framework to clarify the empirical objects motivating the identification analysis, and then describes the Israeli administrative data used to illustrate the theory throughout the paper.

1.1 Normalized Event Studies

This section briefly presents the normalized event-study empirical strategy commonly used in child-penalty applications (e.g., Andresen & Nix, 2022; de la Vega, 2022; Kleven, 2022; Kleven, Landais, Posch, et al., 2019; Kleven, Landais, & Sogaard, 2019; Kleven et al., 2021).

I briefly present notation needed to formulate the normalized event study approach. Consider a finite population of individuals $i \in \{1, \dots, n\}$ observed over a finite set of time periods $t \in \{1, \dots, T\}$. Let $Y_{i,a}$ denote real annual labor-market earnings of individual i at age a, and let D_i denote the age at which individual i has their first child. Define the event time as $E_{i,a} = a - D_i$. Let $G_i \in \{f, m\}$ indicate gender, where f represents female and m male.

The estimation algorithm for the normalized event-study proceeds in three steps. First, outcomes are regressed on event-time indicators and fixed effects separately for each gender. The regression model for gender $g \in \{f, m\}$ is

$$Y_{i,a} = \sum_{e \neq -1} \beta_e^g 1\{E_{i,a} = e\} + \alpha_a^g + \alpha_t^g + u_{i,a}, \tag{1}$$

where $1\{\cdot\}$ denotes an indicator function, α_a^g and α_t^g are age and year fixed effects, respectively, and the superscript g indexes gender-specific coefficients. In the second step, predicted earnings net of event-time coefficients are computed as $\widetilde{Y}_{i,a} = \widehat{\alpha}_a^{G_i} + \widehat{\alpha}_t^{G_i}$. Finally, the esti-

mated event-time coefficients $\widehat{\beta}_e^g$ are normalized by the conditional mean of $\widetilde{Y}_{i,a}$ within gender and event time:

$$\widehat{\theta}_{\text{ES}}(g,e) = \frac{\widehat{\beta}_e^g}{\mathbb{E}_n\left[\widetilde{Y}_a \mid G = g, E_a = e\right]}, \tag{2}$$

where $\mathbb{E}_n[\cdot]$ denotes the sample mean.

Recent work has shown that two-way fixed effects regressions similar to (1) can produce biased estimates in the presence of multiple treatment groups (Borusyak et al., 2024; De Chaisemartin & d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Sun & Abraham, 2021). As Melentyeva and Riedel (2023) argue, this concern applies to the child penalty setting as well. Our focus is not on biases from the estimation procedure itself, but rather on biases arising from the underlying identification assumptions. In the section below I turn to articulating the identification framework that is rationalized from the above normalized event study empirical strategy.

1.2 Data

This section introduces the data used to illustrate the theoretical discussion. Although the arguments developed below are theoretical and generalizable, I illustrate their implications using a specific application to aid with constructing intuition, empirically assess key claims, and document new empirical insights. To that end, I describe the Israeli administrative data used throughout the paper, including its sources, variables, and sample definitions.

The raw dataset covers all Israeli citizens born between 1970 and 2000, matched to their spouses, parents, and children. It was compiled by the Israeli Central Bureau of Statistics (CBS) and integrates data from several administrative sources, including the Population and Immigration Authority's Civil Registry, the Ministry of Education, the Council for Higher Education, and the Israeli Income Tax Authority.

1.2.1. Main Variables. This subsection describes the main variables used in the analysis.

Treatment: Age at birth of first child. Each individual is linked to their biological children. The year of birth of the earliest child is used to define the year of first childbirth. Subtracting the parent's year of birth yields their age at first birth.

Outcome: Earnings. Annual labor market earnings are observed from 1999 to 2020, based on micro-level tax records. Earnings are coded as zero in years with no reported income. All values are expressed in real 2020 New Israeli Shekels (NIS), using the CBS consumer price index.

The analysis below makes use of several additional variables, defined explicitly in Ap-

pendix A.1. These include grandparents' earnings, nationally administrated mathematics test scores called Meitsav, number of credits in high-school subjects, university psychometric entrance test (UPET) scores, years of education and highest education degree. The sample definition also uses ethnicity and religion variables.

1.2.2. Sample Definitions. I make the following restrictions on the main sample. First, the analysis focuses on non-Haredi Jews; individuals identified as Arab or Ultra-Orthodox (Haredi) Jews are excluded, due to systematically different fertility and labor market trajectories (Gould & Lichtinger, 2024; Yakin, 2021). I further restrict the sample to individuals born between 1975 and 1990. Older cohorts are observed only at advanced ages, while younger cohorts are only partially observed through their late twenties. When adding controls, we limit to birth cohorts 1980 and older, reflecting data availability for education variables.

Furthermore, I drop individual-year observations where the individual is less than 20 years old, corresponding to the typical entry into the labor market after high school completion and mandatory army service.[2] I also drop parents who had their first child at ages prior 24 or post 40. Births before age 24 would imply pre-trend diagnostics occur before age 20, where few observations exist, while births after age 40 involve very small sample sizes. Additionally, I keep years only in time window where individuals are reported as alive by the Civil Registry. The dataset used in the main analysis, after the above limitations, consists of 13.7 million individual-year observations, made up of 374 thousand mothers and 320 thousand fathers. Further construction details are provided in Appendix A.2.

1.2.3. Sample Statistics. Appendix Figure G1 reports event studies, specifically estimates of (2), using the Israeli data. Consistent with findings in other countries, the normalized effects for mothers and fathers, $\widehat{\theta}_{\text{ES}}(f,e)$ and $\widehat{\theta}_{\text{ES}}(m,e)$, are very similar before childbirth. After childbirth, a gender gap emerges, with mothers' normalized effects more negative than fathers'. [3] While these patterns mirror the main qualitative features found in other studies, interpreting them causally requires caution, as discussed below.

[2]The legal minimum working age in Israel is 15. However, most individuals complete high school at 18, followed by mandatory army service—two years for women and three years for men. Some individuals, such as religious women, may be exempt from military service but instead perform national service (e.g., in schools or hospitals).

[3]The Israeli results display two patterns that differ somewhat from what is typically observed in other countries. First, the pre-trends for both mothers and fathers are not flat but negative; however, they remain parallel, consistent with the main interpretation of event studies. Second, fathers' normalized effect $\widehat{\theta}_{\text{ES}}(m,e)$ declines over time after childbirth. Despite these differences, the magnitude and persistence of the gender gap are broadly in line with findings from other settings. Finally, my estimates are similar to other literature on child penalty in Israel, e.g., Gould and Lichtinger (2024).

2 The Rationalized Identification Framework

This section develops the identification framework for normalized event studies. I first introduce the potential-outcomes notation and define the target causal estimands. I then show that the identification assumption, rationalized by the pre-treatment validation tests, is equal bias in the violation of normalized parallel trends across genders. I finish with showing that even when this assumption holds, the target causal estimand, the gender gap in normalized effects, is not identifiable when the parallel trends assumption (in levels) is violated. Throughout, for a given gender g I will focus on 2×2 comparisons: treatment group d, control group d', target age a and pre-treatment age $d-1$. Following Melentyeva and Riedel (2023), I set the control group to be the closest-not-yet-treated treatment group, i.e., $d' = a+1$. For example, if $a = 30$ then $d' = 31$, if $a = 31$ then $d' = 32$, and so on. I return to aggregation across treatment groups in Section 5.

2.1 Potential Outcomes

Let $W_{i,a} = 1_{\{a \geq D_i\}}$ denote the treatment status of individual i at age a, where D_i is the age at first childbirth. This definition implies that child penalties are a staggered adoption design, i.e., $W_{i,a-1} = 1 \rightarrow W_{i,a} = 1$. In this design, each group of parents who experience first birth at a given age $D_i = d$ is treated as a distinct treatment group, untreated before d and treated from age d and onward.

Under the stable unit treatment value assumption (SUTVA) (Rubin, 1980), in a staggered adoption design potential outcomes are a function of the timing of treatment (see, e.g., Callaway & Sant'Anna, 2021).[4] Formally, let $Y_{i,a}(d)$ be the potential outcome of individual i at age a if the first childbirth occurs at age d. Let $Y_{i,a}(\infty)$ denote the potential outcome if i never has a child. Observed outcomes are linked to potential outcomes by the consistency assumption: $Y_{i,a} = Y_{i,a}(D_i)$.

2.2 Causal Estimands

This subsection defines the causal estimands of interest in the child penalty context. I begin with estimands for a single treatment group and a specific gender, and then consider

[4] This requires that "age at first birth" satisfies the assumption known as treatment variation irrelevance (VanderWeele, 2009), one of the two elements in SUTVA. For example, for a mother who gave birth to her first child at age 25, a counterfactual scenario in which she delays childbirth to age 35 could arise through many distinct causal mechanisms, such as divorce, health issues, or career disruptions. If these different versions of the treatment yield different counterfactual earnings, the potential outcome $Y_{i,a}(d')$ is ill-defined. Since this issue is beyond the scope of the current paper, we abstract from it and assume well-defined potential outcomes throughout.

estimands of differences between genders. In Section 5.1 I discuss causal estimands which aggregate multiple treatment groups.

Let $APO(g, d, d', a) = \mathbb{E}[Y_a(d') \mid G = g, D = d]$, denote the average potential outcome (APO) at age a for individuals of gender g who had their first child at age d, had they instead had their first child at age d'. Next, define the average treatment effect (ATE) for gender g, treatment group d at age a as $ATE(g, d, a) = APO(g, d, d, a) - APO(g, d, \infty, a)$, where $APO(g, d, \infty, a)$ denotes the average potential outcome for individuals from treatment group d and gender g in the counterfactual of never having children. To mirror the event study estimator $\widehat{\theta}_{\text{ES}}(g, e)$ in (2), we define the causal estimand

$$\theta(g, d, a) = \frac{ATE(g, d, a)}{APO(g, d, \infty, a)}.$$

$\theta(g, d, a)$ captures the proportional earnings loss from childbirth at age a, relative to the counterfactual of never giving birth.[5] This provides a normalized measure of the effect of parenthood on earnings age a for individuals of gender g from treatment group d.[6]

The term "child penalty" is frequently used to describe the differential impact of parenthood on labor market outcomes between women and men. Several causal estimands can capture this gender gap. A natural starting point is $ATE(f, d, a) - ATE(m, d, a)$, which reflects the level difference in the impact of parenthood on earnings between females and males. However, since earnings levels may differ by gender even in the absence of children, researchers may prefer a normalized comparison. One such alternative is $\theta(f, d, a) - \theta(m, d, a)$, which captures the gender gap in relative earnings losses from parenthood, that is, the gender gap in normalized effects. Since the normalized event study approach (Section 1.1) compares $\widehat{\theta}_{\text{ES}}(g, e)$ in (2) across gender, $\theta(f, d, a) - \theta(m, d, a)$ represents the causal estimand implicitly targeted in that empirical strategy.

2.3 Descriptive Estimands

By a descriptive estimand, I mean an expectation defined solely in terms of observed outcomes and covariates, without involving potential outcomes beyond the realized outcome Y (as in Abadie et al., 2020). The following three descriptive estimands—used below to

[5] The normalized average treatment effect is similar to estimands studied in the vaccine efficacy literature (see, e.g., Orenstein et al., 1985), and is also related to target estimands in the excess mortality literature (see, e.g., Msemburi et al., 2023).

[6] This interpretation aligns with the estimand targeted in child penalty studies, i.e., $\widehat{\theta}_{\text{ES}}(g, e)$ in (2). For example, Kleven, Landais, and Sogaard (2019, p. 188) describe their child penalty estimator (P_t^g in their notation) as "the year-t effect of children as a percentage of the counterfactual outcome absent children," where t corresponds to event time e in my notation. Similar quotes can be found in other papers that use the normalized event study strategy.

identify causal estimands and to construct validation tests—illustrate how DID can be used to construct the counterfactual APO, ATE and θ.

$$\begin{aligned}
\delta_{\text{APO}}(g,d,d',a) &= \mathbb{E}[Y_{d-1} \mid G=g, D=d] + \mathbb{E}[Y_a - Y_{d-1} \mid G=g, D=d'], \\
\delta_{\text{ATE}}(g,d,d',a) &= \mathbb{E}[Y_a \mid G=g, D=d] - \delta_{\text{APO}}(g,d,d',a), \\
\delta_{\theta}(g,d,d',a) &= \frac{\delta_{\text{ATE}}(g,d,d',a)}{\delta_{\text{APO}}(g,d,d',a)}.
\end{aligned} \tag{3}$$

In δ_{APO}, the first term provides the pre-treatment level from the treated group, and the second term adds the trend from the control group. Hence, δ_{APO} is how DID constructs the counterfactual APO for the treatment group. δ_{ATE} is the conventional DID four expectations and three differences estimand. In our context, it is how DID constructs the effect of parenthood on earnings in levels. δ_θ is the ratio of these two, and hence how DID can be used to construct the normalized effect.

2.4 The Identification Assumption

I now turn to derive the identification assumption that is implicitly maintained in normalized event studies, as inferred from the empirical validation test commonly used in applied work. Before presenting the result, I introduce the no anticipation identification assumption and introduce needed notation.

The no anticipation assumption requires that potential outcomes before childbirth are the same under the observed treatment path and the counterfactual of never giving birth (Abbring & Van den Berg, 2003).[7] Formally,

Assumption NA (No Anticipation)**.** For gender g, treatment age d, and target age $a < d$, $APO(g,d,d,a) = APO(g,d,\infty,a)$.

Next, define the difference in counterfactual earnings trends from age $d-1$ to a between treatment group d and control group d' as:

$$\begin{aligned}
\gamma_{\text{PT}}(g,d,d',a) = {} & APO(g,d,\infty,a) - APO(g,d,\infty,d-1) \\
& - \left[APO(g,d',\infty,a) - APO(g,d',\infty,d-1)\right].
\end{aligned} \tag{4}$$

The statement $\gamma_{\text{PT}}(g,d,d',a) = 0$, i.e., that counterfactual trends are equal across treatment and control, is often called the parallel trends identification assumption in the DID

[7] If outcomes at age $d-1$ are affected by anticipatory behavior, the no anticipation assumption can instead be imposed at earlier ages (e.g., $d-2$ or $d-3$), shifting the baseline period for the treated group. Such concerns may also motivate shifting the closest-not-yet control group may to later ages (e.g., $a+2$ or $a+3$).

framework. Formally, it can be written as

Assumption DID-PT (Parallel Trends)**.** For gender g, treatment group d, control group d', and target age a, $\gamma_{\text{PT}}(g, d, d', a) = 0$.

To derive the identification assumption underlying the normalized event-study approach, I start from the validation test that is typically conducted in practice. Empirically, researchers examine whether, prior to childbirth, the gender gap in normalized effects is zero, that is, whether the estimates satisfy $e < 0 : \widehat{\theta}_{\text{ES}}(f, e) = \widehat{\theta}_{\text{ES}}(m, e)$. The question is what identifying assumption such a test actually validates. To backward-engineer this assumption, I restate the test in a 2×2 comparison and replace estimators with descriptive estimands. The 2×2 equivalent can be written as follows: For treatment group d, control group d', and pre-treatment age $a < d$,

$$a < d < d' : \delta_\theta(f, d, d', a) - \delta_\theta(m, d, d', a) = 0. \tag{5}$$

That is, in a 2×2 where the target age is pre-treatment, difference the DID equivalent of θ, denoted δ_θ from Section 2.3, across gender, and require this difference be equal to zero. The following result states the identification assumption that is implicitly tested by this restriction.

Proposition 1. *Consider a* 2×2 *with treatment group* d*, control group* $d' > d$*, and pre-treatment ages* $d - 1$ *and* $a < d - 1$*. If Assumption NA holds then*

$$\delta_\theta(f, d, d', a) = \delta_\theta(m, d, d', a) \quad \iff \quad \frac{\gamma_{\text{PT}}(f, d, d', a)}{APO(f, d, \infty, a)} = \frac{\gamma_{\text{PT}}(m, d, d', a)}{APO(m, d, \infty, a)}.$$

The proof is presented in Appendix B. Proposition 1 characterizes the restriction on potential outcomes that is implied, in pre-treatment periods, when the gender gap in δ_θ is equal to zero. I now state this restriction formally as a new identification assumption:

Assumption NTD-PT (Equal Difference in Normalized Trends)**.** For treatment group d, control group d' and target age a,

$$\frac{\gamma_{\text{PT}}(f, d, d', a)}{APO(f, d, \infty, a)} = \frac{\gamma_{\text{PT}}(m, d, d', a)}{APO(m, d, \infty, a)}.$$

Assumption NTD-PT states that the violations of parallel trends, once normalized by the counterfactual APO, are equal across genders. Given Proposition 1, Assumption NTD-PT can be interpreted as the implicit identification assumption in normalized event-study

strategies as in Section 1.1. Going forward, I refer to the identification framework based on this assumption as the Normalized Triple Differences (NTD) framework.

2.5 The (Un-)Identification Result

I now present the first main result of the paper: even when the identification assumption implicit in normalized event studies holds, the target causal estimand, the gender gap in normalized effects $\theta(f,d,a)-\theta(m,d,a)$, is not identifiable. The following explicitly characterizes this result.

Theorem 1. *Consider a 2×2 with treatment group d, post-treatment age a, and control group d' such that $d'>a\geq d$, and assume Assumptions NA and NTD-PT hold. Then*

$$\delta_\theta(f,d,d',a)-\delta_\theta(m,d,d',a)=\text{Bias}(d,d',a)\big[\theta(f,d,a)-\theta(m,d,a)\big],$$

where

$$\text{Bias}(d,d',a)=\frac{APO(f,d,\infty,a)}{APO(f,d,\infty,a)-\gamma_{\text{PT}}(f,d,d',a)}=\frac{APO(m,d,\infty,a)}{APO(m,d,\infty,a)-\gamma_{\text{PT}}(m,d,d',a)}.$$

Theorem 1 shows that the normalized event-study empirical strategy discussed in Section 1.1 does not recover the causal estimands it is often interpreted as capturing, since the NTD framework allows violations of Assumption DID-PT.[8] I return to the validity of Assumption DID-PT in Section 3.1.

For a sketch of the proof, consider the gender gap in δ_θ. Suppressing non-gender inputs for brevity (e.g., $APO(m,d,\infty,a)=APO(m)$), we obtain[9]

$$\delta_\theta(f)-\delta_\theta(m)=\theta(f)\,Bias_{\text{Mult}}(f)-\theta(m)\,Bias_{\text{Mult}}(m)+Bias_{\text{Add}}(f)-Bias_{\text{Add}}(m),$$
$$Bias_{\text{Mult}}(g)=\frac{APO(g)}{APO(g)-\gamma_{\text{PT}}(g)},\quad Bias_{\text{Add}}(g)=\frac{\gamma_{\text{PT}}(g)}{APO(g)-\gamma_{\text{PT}}(g)}.$$

Identification is achieved if the additive bias $Bias_{\text{Add}}(g)$ is constant across gender, and the multiplicative bias $Bias_{\text{Mult}}(g)$ is constant across gender and equal to one. Under these requirements, $\delta_\theta(f)-\delta_\theta(m)=\theta(f)-\theta(m)$ is achieved. However, under Assumption NTD-PT, while the additive and multiplicative biases are indeed constant by gender, they are not

[8] Lemma 2 in Appendix B shows that $P(d,a)=[ATE(f,d,a)-ATE(m,d,a)]/APO(f,d,\infty,a)$, the 2×2 analogue of the estimator that Kleven, Landais, and Sogaard (2019) term the child penalty (P_t in their notation), is also not identifiable under NTD when parallel trends in levels fail. I focus instead on the gender gap in normalized effects, $\theta(f,d,a)-\theta(m,d,a)$, since most papers using the normalized event-study approach in Section 1.1 discuss the gender gap in $\widehat{\theta}_{\text{ES}}$ from (2) as their main result.

[9] The result follows from the proof of Theorem 1.

guaranteed to be equal to one, and hence bias remains as in Theorem 1.

To summarize, the discussion above shows that the widely used normalized event study framework relies implicitly on an identification assumption that does not allow identification of the gender gap in normalized effects. To achieve identification of the effect of parenthood on labor market outcomes, the rest of the paper branches into two distinct paths. The first is to switch frameworks: specifically, return to the level-based difference-in-differences (DID) or triple-differences (TD) designs, and ask whether their identifying assumptions are valid in the child penalty application. The second is to keep using the NTD framework, but consider alternatives that allow identification. Namely, either redefine the target estimand to one that is identified under the existing assumptions, or add additional assumptions that allow identification of the original causal estimand. I start with the first branch which examines DID and TD as viable alternative frameworks, and then continue with the second branch which explores alternatives within NTD.

3 Alternative Frameworks

Motivated by the result that the NTD framework fails to identify child penalties, this section examines whether DID and TD provide viable alternatives. It is well known that if their identification assumptions hold, these frameworks identify causal estimands (e.g., Angrist & Pischke, 2009). The main question is therefore whether those assumptions are plausible in the child-penalty context. I begin with the DID framework and draw on insights from economic models of human capital accumulation to highlight a mechanism that likely violates the parallel-trends assumption for certain comparisons. For TD, I derive a new decomposition that links the gender gap in parallel-trends violations to counterfactual gender inequality, and use the new decomposition to discuss the bias in the TD framework during post-treatment periods. As the focus of this section is validating identification assumptions, I conclude by discussing, both theoretically and empirically, validation tests based on pre-treatment periods, where the control group changes with the target age. A possible modification is to include covariates in the identification assumptions; Appendix E presents DID and TD identification and estimation with covariates building on Callaway and Sant'Anna (2021) and Leventer (2025). Appendix F discusses DID and TD with logarithmic functional form.

3.1 DID and Selection on Treatment

In this subsection I argue the parallel-trends assumption (Assumption DID-PT) is unlikely to hold because fertility timing, the treatment variable, is endogenous with respect to potential

earnings.[10] Specifically, I argue that differences in human capital among treatment groups likely cause DID to overstate the effect of parenthood on earnings. I then provide empirical evidence in line with positive selection on treatment; later treatment groups have higher observed human and social capital.

3.1.1 Theoretical Argument. A large body of theoretical and empirical work argues that earnings trajectories reflect underlying differences in human and social capital (e.g., Ben-Porath, 1967; Cunha & Heckman, 2007; Heckman, 1976; Heckman & Mosso, 2014), and that the timing of childbirth responds to these same factors (e.g., Becker et al., 1990; De La Croix & Doepke, 2003; Geronimus & Korenman, 1992). These two strands of the literature are combined in life-cycle models of fertility and labor (e.g., Adda et al., 2017; Blackburn et al., 1993; Eckstein et al., 2019; Francesconi, 2002; Jakobsen et al., 2022; Keane & Wolpin, 2010; Moffitt, 1984).[11] In the considered context, such models have straightforward implications for the parallel trends assumption (Assumption DID-PT). First, individuals with higher labor-market ability invest more heavily in human capital and, as a result, delay fertility. That is, later treatment groups are positively selected on ability relative to earlier treatment groups. Second, considering life-cycle profiles of $Y_{i,a}(\infty)$—the potential earnings under the counterfactual without children—individuals with higher labor-market ability invest more in human capital and hence have steeper trajectories, particularly early in their careers (e.g., ages 25–35). For early treatment groups—for instance, $D = 25$—the relevant post-treatment ages are the late 20s and early 30s. At these ages, the not-yet-treated control groups (e.g., $D = 2931$) consist of individuals with higher human capital who would, even absent children, be on steeper earnings trajectories. Consequently, the parallel-trends violation is expected to be negative for early treatment groups at later ages.

Given $\gamma_{\text{PT}} < 0$, DID inflates the counterfactual earnings imputed for early-treated parents. Formally, under Assumption NA, $\delta_{\text{APO}} = APO - \gamma_{\text{PT}} > APO$ (Lemma 1 in Appendix B), where I abuse notation by omitting function inputs for brevity. This upward shift in counterfactual earnings implies that the DID estimand of the treatment effect is biased downward

[10] The no anticipation assumption (Assumption NA) is also unlikely to hold. The literature shows that households factor future fertility into career decisions (e.g., Angrist & Evans, 1996; Attanasio et al., 2008; Blundell et al., 2018; Cristia, 2008; Doepke & Kindermann, 2019; Gronau, 1977; Hazan & Zoabi, 2015). This paper focuses on violations of parallel trends. Developing analogous strategies to cope with violations of the no anticipation assumption is an important direction for future research.

[11] Adda et al. (2017) emphasize that family preferences, rather than initial ability differences, drive occupational sorting differences between early and late fertility mothers. Yet, because fertility preferences shape early-life career and education choices, observed treatment groups will still diverge in their earnings trajectories even in the counterfactual where they do not ultimately have children. Thus, a positive correlation between delayed fertility and human capital investment emerges through family preferences and expected fertility rather than innate ability.

relative to the true ATE: $\delta_{\text{ATE}} = ATE + \gamma_{\text{PT}} < ATE$ (Lemma 1 in Appendix B). Because $|\delta_{\text{APO}}| > |\delta_{\text{ATE}}|$ it follows that $\delta_\theta < \theta$. In words, the above economic reasoning suggests that, due to positive selection of later treatment groups on human capital, DID overstates counterfactual earnings for early-treated parents, and in turn overstates the causal effect of parenthood on earnings in both levels and ratios.

3.1.2 Empirical Evidence. Figure 1 plots several early-life indicators of human and social capital against parents' age at first birth, separately by gender. Variable definitions are provided in Appendix A.1, while Appendix A.2 details data availability constraints that determine the sample used for each measure. The figure shows that parents who have their first child around age 30, compared to those who give birth earlier, come from families with higher-earning and more educated parents, achieve higher scores on national mathematics exams in primary and middle school, and are more likely to take advanced tracks in high school. Melentyeva and Riedel (2023) document similar evidence for grandparents' education in Germany. Jensen et al. (2024), Appendix Figure G3, and Appendix Figure G4 show similar evidence for Denmark, the United States, and Israel when considering final educational attainment. For Israel, Appendix Figure G4 documents similar patterns for additional later-life educational outcomes, such as probability of taking and, conditional on taking, the average score in the UPET.[12]

Taken together, the evidence suggests there exists positive selection on human and social capital into later fertility.[13] The findings further suggest that selection is strongest between ages 20 and 30, then flattens or declines somewhat beyond 30. Selection patterns for fathers peak slightly later than for mothers, consistent with the two-year average spousal age gap and assortative matching.

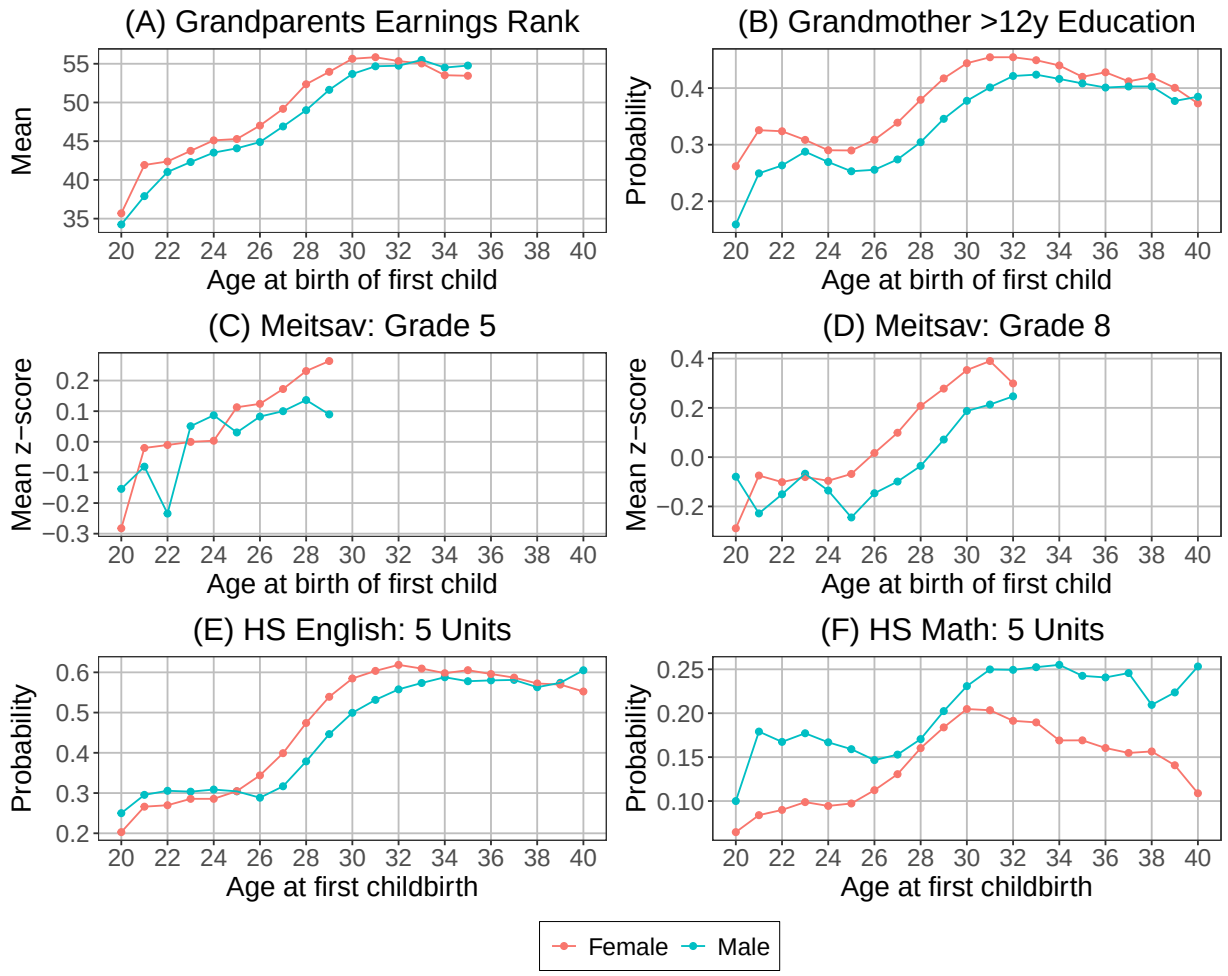
3.2 TD and Counterfactual Gender Inequality

TD is a viable alternative as it offers two potential advantages. First, because it is defined in levels, it avoids the normalization bias identified in Theorem 1. Second, by differencing across genders, it may offset the selection bias arising from fertility timing that affects within-gender comparisons, as discussed in the previous subsection. Formally, the TD framework relaxes the parallel-trends assumption within gender but requires that its violations are equal across genders. This assumption can be written as:

[12]These findings relate to the literature on the causal effect of education on fertility timing, in particular the age at first birth (e.g., Black et al., 2008; McCrary & Royer, 2011).

[13]Prior studies document a link between delayed childbearing and human capital observables (e.g., Buckles, 2008), but rely on post-treatment variables such as career, earnings or final education. In contrast, I use early-life measures, minimizing concerns about reverse causality.

Figure 1: Selection and Age Birth of First Child



Notes: The figure presents means of ranked grandparents earnings when the parents were aged 5-10 (Panel A), probability of grandmothers education being greater than 12 years (Panel B), means of normalized Meitsav mathematics test scores in grades 5 and 8 (Panels C and D, respectively), and probabilities of taking 5-unit tracks in English and mathematics in high school (Panels E and F, respectively), by age at birth of first child (x-axis) and gender (colors). Meitsav is a national test administered by the Ministry of Education. 5 units is the highest number which can be taken. The sample changes by the considered variable due to different data-constraints, discussed in detail in Appendix A.

Assumption TD-PT (Equal Difference in Trends)**.** For treatment group d, control group d', and target age a, $\gamma_{\text{PT}}(f, d, d', a) = \gamma_{\text{PT}}(m, d, d', a)$.

To assess the validity of this assumption, I derive a decomposition that links the gender gap in γ_{PT} to life-cycle counterfactual gender earnings ratios. I then turn to the data, first documenting pre-birth gender earnings ratios, which under no anticipation identify the counterfactual gender earning ratios, and then conduct an empirical exercise suggesting that

for early treatment groups at later ages, the TD framework likely understates the gender gap in the effects of parenthood.

3.2.1 Theoretical Argument. Let $\rho(d, d', a) = \frac{APO(f,d,d',a)}{APO(m,d,d',a)}$ denote gender inequality in earnings for treatment group d at age a, in the counterfactual had they instead given birth at d'. Substituting $\rho(d, \infty, a)$ into the expression for the female parallel-trends violation and differencing across genders yields the following decomposition:

$$\begin{aligned}
&\gamma_{\text{PT}}(f, d, d', a) - \gamma_{\text{PT}}(m, d, d', a) \\
&= \big(\rho(d, \infty, a) - 1\big) APO(m, d, \infty, a) - \big(\rho(d, \infty, d-1) - 1\big) APO(m, d, \infty, d-1) \\
&- \Big[\big(\rho(d', \infty, a) - 1\big) APO(m, d', \infty, a) - \big(\rho(d', \infty, d-1) - 1\big) APO(m, d', \infty, d-1)\Big]. \quad (6)
\end{aligned}$$

Expression (6), which to my knowledge is new to the literature, clarifies how the TD identification assumption is tied to assumptions about $\rho(d, \infty, a)$.

Two cases illustrate the logic. If $\rho(d, \infty, a)$ is constant and equal to one then $\gamma_{\text{PT}}(f, d, d', a) = \gamma_{\text{PT}}(m, d, d', a)$, that is Assumption TD-PT holds. Assumption TD-PT can still hold with $\rho(d, \infty, a) \neq 1$, provided that the four terms on the right-hand side of (6) cancel out. Alternatively, if $\rho(d, \infty, a) = \rho \neq 1$ is constant but not equal to one, then $\gamma_{\text{PT}}(f, d, d', a) = \rho \times \gamma_{\text{PT}}(m, d, d', a)$, that is Assumption TD-PT fails unless parallel trends holds.[14]

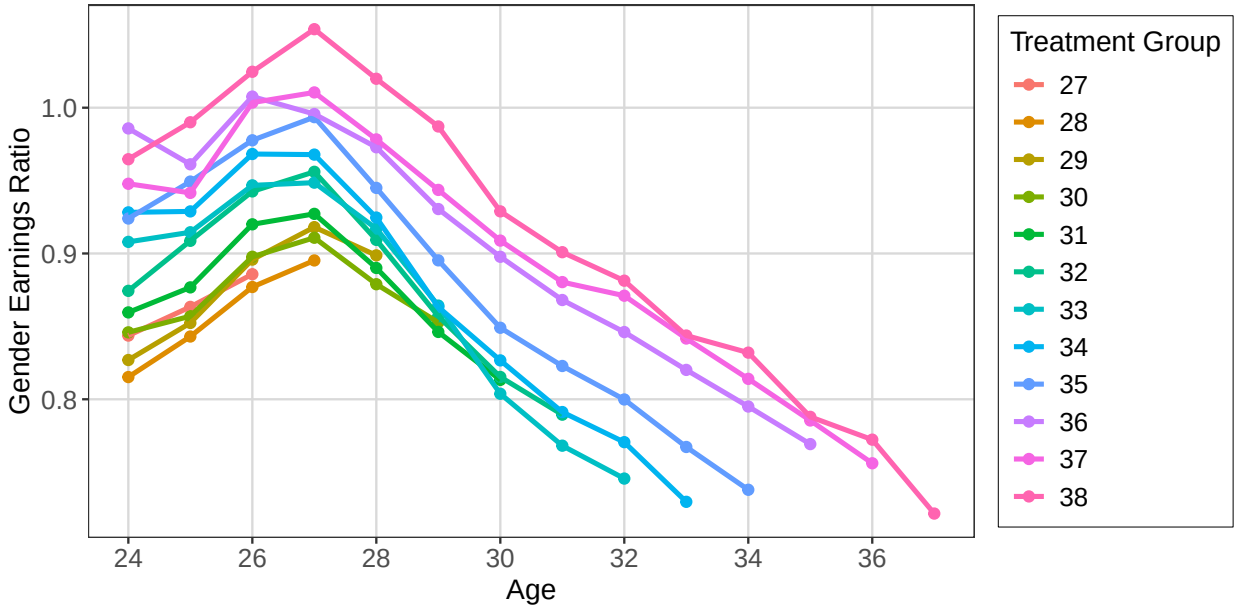
Selection on treatment, as discussed in Section 3.1, suggests that later treatment groups have higher APO values at later ages due. However, theory offers little guidance on how $\rho(d, \infty, a)$ evolves across groups and ages. Since $\rho(d, \infty, a) = 1$ for all d, a is implausible in most contexts, as gender inequality is likely present even in the absence of children, TD is unlikely to hold exactly. Yet the direction and magnitude of the violation cannot be determined without information on $\rho(d, \infty, a)$. Next, I turn to the data to document empirical patterns of $\rho(d, \infty, a)$ in Israel, and illustrate a general strategy—combining the decomposition with pre-treatment data and the selection-on-treatment argument—to assess the sign of TD violations in applications.

3.2.2 Empirical Evidence. To examine how counterfactual gender inequality $\rho(d, \infty, a)$ evolves across the life cycle and treatment groups, I proceed as follows. For each treatment age d, I compute mean earnings at every age $a < d$ and form the female-to-male observed earning ratio $\widehat{\rho}(d, \infty, a)$. Under the no-anticipation assumption (Assumption NA), these pre-childbirth earnings identify the relevant APOs, so $\widehat{\rho}(d, \infty, a)$ identifies $\rho(d, \infty, a)$.

Figure 2 plots $\widehat{\rho}(d, \infty, a)$ for each treatment group $d \in [27, 38]$. Two descriptive patterns

[14] A similar decomposition can be written for Assumption NTD-PT. Such a decomposition shows that if $\rho(d, \infty, a)$ is constant, even if not at one, the assumption holds.

Figure 2: Observed Gender Earnings Ratios Before Childbirth



Notes: The figure presents the ratio of mean female to mean male earnings (y-axis) for pre-treatment ages separately by treatment group (colors). The sample is restricted to treatment groups $D \in [27, 38]$. The dashed horizontal line indicates gender equality in earnings. Under the no-anticipation assumption (Assumption NA), pre-birth earnings equal the *APO*s under the never giving birth counterfactual, and hence the series provide empirical estimates of the counterfactual gender inequality measure $\rho(d, \infty, a)$.

emerge. First, within treatment group the gender earnings ratio before childbirth follows a similar life-cycle pattern: rising up to age 27, and falling from age 28, producing an inverted U-shape. Second, at any given age, parents who delay their first birth generally exhibit higher ratios than earlier-childbearing parents.

3.2.3 Forming Predictions on Violation Sign. The decomposition in (6) together with the the life-cycle pattern of ρ presented above, can be used to provide suggestive evidence on the sign of bias in TD arising from violations of its identification assumption. Since this exercise is not directly related to the main results on NTD, I defer it to Appendix C. Implementing the exercise using the Israeli data, I find that TD likely understates child penalties for early treatment groups.

3.3 Validation Tests

This subsection discusses the need to conduct disaggregated validation tests by treatment–control pairs when the control group changes across post-treatment periods. I then implement these tests using Israeli data and find that DID estimates in pre-treatment periods grow

in magnitude as the age gap between treatment and control widens, while TD and NTD pre-trends remain small and stable across control groups for mid-range treatment groups.

3.3.1. Theory. In DID, TD and NTD, the conventional validation approach is to test for zero differences in pre-treatment periods, commonly referred to as "pre-trends" testing (Autor, 2003; Roth, 2022). As discussed in Section 2, I consider 2×2 comparisons with the closest not-yet-treated group serving as the control group. Hence, the number of control groups is determined by the set of event times the researcher analyzes in post-treatment periods. For example, estimating the child penalty for event times $e = 0, \ldots, 5$ requires six distinct control groups. Since Assumptions DID-PT, TD-PT and NTD-PT must hold separately for each treatment–control pair, pre-trends testing should also be performed separately for each treatment-control pair.[15]

An additional consideration concerns aggregation. Event-study plots, such as those discussed in Section 1.1, typically display pre-trends aggregated across all treatment and control groups. However, aggregation across pre-treatment periods does not validate the identification assumption, which is made separately for each treatment–control pair. Evaluating disaggregated pre-trends is therefore essential.

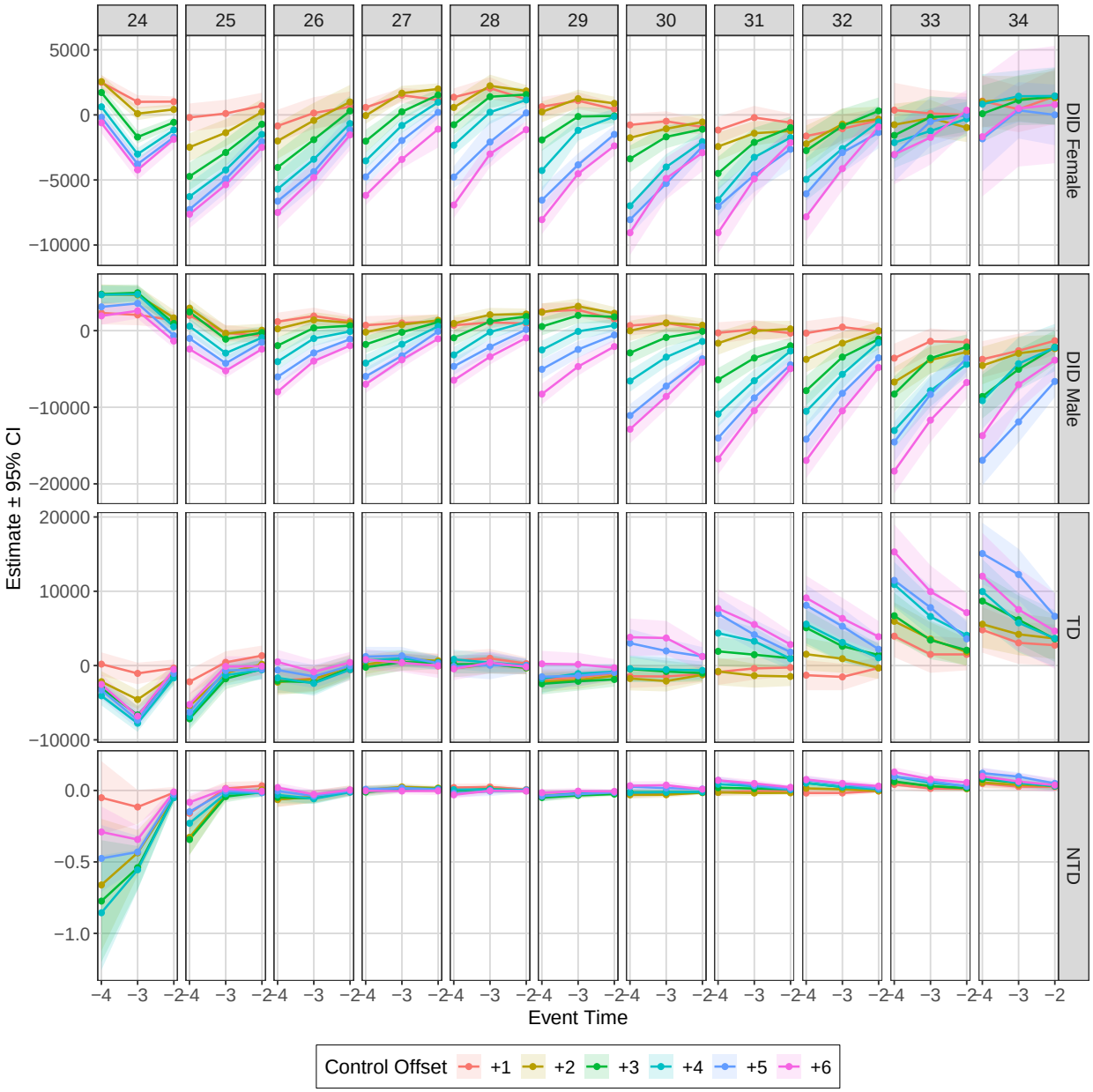
3.3.2. Empirical Evidence. Following the above discussion, Figure 3 reports pre-treatment validation tests for treatment–control pairs across frameworks in the Israeli application. In the results below, I report estimates up to five years post-treatment, and hence for each treatment group d the figure reports results for control groups $d' - d = 1, ..., 6$.[16] I include validation tests for NTD because in the next section I develop alternative frameworks that build on NTD. Estimators replace population expectations with sample means, and use influence functions (IF) based estimators for standard errors (SEs); See Appendix D for explicit formulas.

Three patterns stand out. First, ATE estimates, shown in the top two rows, increase in magnitude as the treatment–control age gap widens. For example, comparing treatment group $D = 25$ to control group $D = 26$ (+1 in the figure) shows small economic difference and confidence intervals that include zero, while $D = 25$ vs. $D = 30$ diverges sharply. This holds across almost all treatment groups and for both mothers and fathers. Second, TD

[15] This issue is not specific to the closest-not-yet-treated control group assignment method. Whenever the control group consists of not-yet-treated individuals in a staggered adoption design, the composition of the control group potentially changes at each post-treatment event-time. Therefore, pre-trends testing should be performed separately for each treatment–control pairing.

[16] As discussed in Section 1.2 I do not include earnings data prior age 20. To ensure four pre-treatment years, the youngest treatment group is $D = 24$. I also do not use treatment groups post age 40. Hence to ensure 5 post-treatment years, the oldest treatment group is $D = 34$.

Figure 3: Validation Tests by Treatment-Control Pairs



Notes: The figure presents estimated pre-treatment differences by framework, with treatment groups in columns and control groups in colors. Rows correspond to DID for females, DID for males, TD, and NTD. DID shows differences in mean earnings trends within gender, TD differences these across genders, and NTD differences normalized effects across genders. Control groups correspond to those used for target ages from first childbirth up to five years after, in two-by-two designs with the closest not-yet-treated group as controls. For example, for treatment group $D = 24$ (left-most column), control group $+1$ corresponds to $D = 25$, $+2$ to $D = 26$, and so on. Estimates are shown for event times $e = -4, -3, -2$, relative to age at first childbirth.

and NTD exhibit larger violations for later treatment groups and smaller ones for mid-range groups ($D = 26, \ldots, 30$). This possibly reflects widening counterfactual gender inequality at older ages (Section 3.2), since pre-treatment periods correspond to the early twenties for early-treated groups and to the early thirties for later-treated groups. Finally, TD and NTD pre-treatment estimates appear smaller in magnitude than in DID; however, such differences are difficult to interpret without benchmarking against the estimated effects.

4 Alternatives Within NTD

Above we examined alternative frameworks to identify child penalties, following the result that NTD does not identify the gender gap in normalized effects. This section takes a different route and examines what can be identified when the NTD framework itself holds. The main result is that NTD allows identification of an alternative estimand, the effect of parenthood on gender inequality, without additional assumptions. I then show that, by adding a null-effect assumption for fathers, a bias-corrected formula can recover the original normalized effect estimand. Each approach is first developed theoretically and then illustrated empirically using the Israeli data. Appendix D discusses estimators and influence-functions used for inference.

4.1 Changing the Estimand

This subsection shows that under the NTD framework, the effect of parenthood on the gender earnings ratio is identifiable. I first present the theoretical result, then discuss its relationship to the (unidentified) gender gap in normalized effect estimand, and finish with documenting their differences empirically.

4.1.1. Theory. Theorem 1 showed that under the NTD framework the gender gap in normalized effects is not identified. The next result shows, however, that it does identify a closely related and policy-relevant quantity: the change in the female–male earnings ratio between the observed treatment and the counterfactual of no childbirth. Recall that $\rho(d, d', a) = \frac{APO(f,d,d',a)}{APO(m,d,d',a)}$ gender inequality in earnings for treatment group d at age a, in the counterfactual had they instead had their first child at d'. Let $\Delta\rho(d, a) = \rho(d, d, a) - \rho(d, \infty, a)$.

Theorem 2. *Assume the same setup as in Theorem 1. If Assumption NTD-PT holds, then*

$$\Delta\rho(d, a) = \frac{\mathbb{E}[Y_a \mid G = f, D = d]}{\mathbb{E}[Y_a \mid G = m, D = d]} - \frac{\delta_{\text{APO}}(f, d, d', a)}{\delta_{\text{APO}}(m, d, d', a)}.$$

The proof is provided in Appendix B. Sketching the proof, the observed earnings ratio $\rho(d, d, a)$ is identified directly from observed data by consistency, while the counterfactual ratio $\rho(d, \infty, a)$ is identified via the gender ratio of δ_{APO} under Assumption NTD-PT. The intuition behind this result is that NTD assumes normalized parallel-trends violations are equal across genders, that is, it makes the identifying assumption in ratios, and hence taking ratios across genders cancels out the bias. The result in Theorem 2 can be viewed as a recommendation for researchers that study child penalties under Assumption NTD-PT to estimate the descriptive estimand on the right-hand side of Theorem 2.

Constructing an estimator follows directly from the identification theorem. Replace the population expectations in $\frac{\mathbb{E}[Y_a|G=f,D=d]}{\mathbb{E}[Y_a|G=m,D=d]}$ with sample means to estimate $\widehat{\rho}(d, d, a)$. Similarly, replace the population expectations in (3) with sample means to estimate $\widehat{\delta}_{\text{APO}}(g, d, d', a)$, and compute $\widehat{\rho}(d, \infty, a) = \frac{\widehat{\delta}_{\text{APO}}(f,d,d',a)}{\widehat{\delta}_{\text{APO}}(m,d,d',a)}$. The estimator of the causal estimand in Theorem 2 is then given by $\widehat{\Delta}\rho(d, a) = \widehat{\rho}(d, d, a) - \widehat{\rho}(d, \infty, a)$. For inference, I compute clustered standard errors based on the influence function of this estimator, discussed in Appendix D. There are three main benefits to the influence-function-based inference: (i) simple to calculate, as all estimators are combinations of simple means; (ii) allow the construction of cluster-robust standard errors, which are important in this application; and (iii) are analytical and thus much faster to compute compared to bootstrapped standard errors.

Next, I discuss the relationship between the gender gap in normalized effects estimand and the new effect of parenthood on gender inequality estimand. Their relationship can be expressed as:[17]

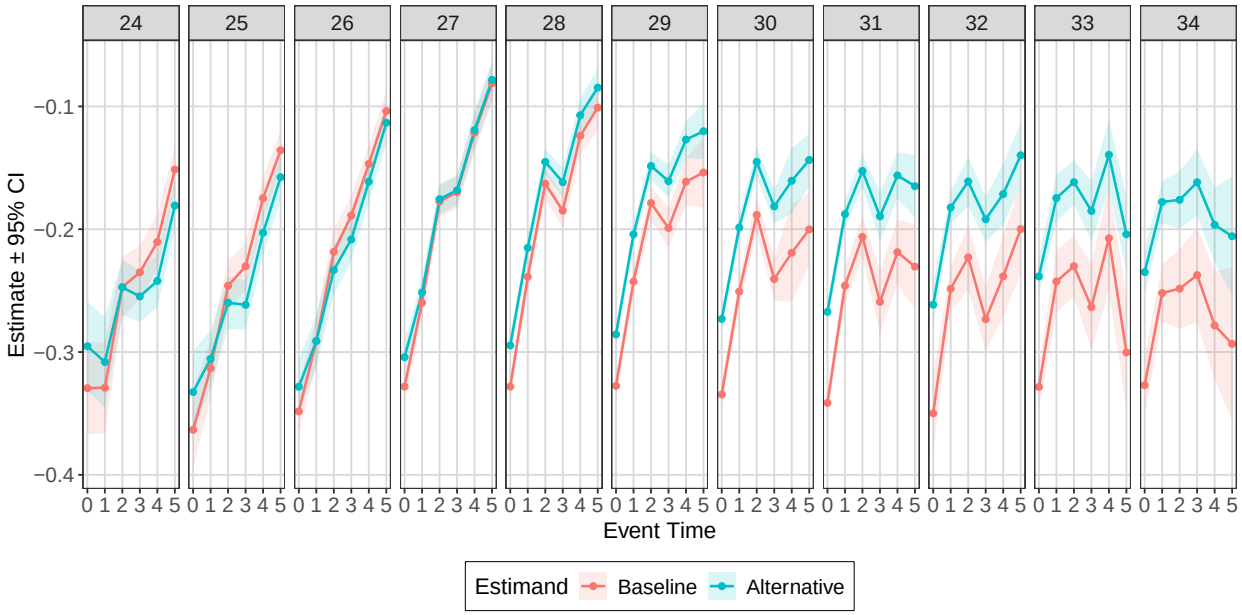
$$\Delta\rho(d, a) = \rho(d, \infty, a) \left[\frac{\theta(f, d, a) - \theta(m, d, a)}{1 + \theta(m, d, a)} \right]. \tag{7}$$

Hence, the new estimand $\Delta\rho(d, a)$ can be written as a rescaled version of the unidentified baseline estimand $\theta(f, d, a) - \theta(m, d, a)$, where the scaling factor depends on the counterfactual gender earnings ratio ($\rho(d, \infty, a)$) and men's normalized effect ($\theta(m, d, a)$). When $\rho(d, \infty, a) < 1$, as suggested by empirical evidence at later ages (Section 3.2), and male effects are close to zero or positive, the new estimand will be smaller in absolute value than the baseline estimand.

5.1.2. Empirical Application. Before presenting the results, recall the validation tests in Section 3.3. These tests suggested that, in the Israeli data, violations of Assumption NTD-PT were smaller and less systematic for treatment groups 26–30, while deviations were more pronounced for earlier and later groups. Accordingly, the discussion below focuses on treatment groups $D \in [26, 30]$, for which the NTD identifying assumption appears most

[17] The derivation of this expression is provided in the proof of Theorem 2.

Figure 4: Comparing Baseline to Alternative Estimand for NTD



Notes: The figure presents estimates for the NTD framework by treatment group (columns) and by estimand (color). The x-axis represents the distance in age from the treatment, age at first childbirth. "Baseline", in red, refers to the gender gap in normalized effects, as discussed in Section 2. "Alternative", in blue, refers to the effect of parenthood on the gender earnings ratio, as discussed in Section 4.1.

plausible. For completeness, Figure 4 presents the results for all treatment groups in the data, $D \in [24, 34]$. The baseline (biased) NTD, in red, estimates $\delta_\theta(f, d, d', a) - \delta_\theta(m, d, d', a)$, and the alternative (unbiased) NTD, in blue, estimates $\frac{\mathbb{E}[Y_a|G=f,D=d]}{\mathbb{E}[Y_a|G=m,D=d]} - \frac{\delta_{\text{APO}}(f,d,d',a)}{\delta_{\text{APO}}(m,d,d',a)}$. All estimators replace population expectations with their sample analogs.

The figure shows the following patterns. Across treatment groups $D \in [26, 30]$, both series have a similar trajectory: a sharp drop at childbirth followed by gradual recovery, with a faster recovery for earlier treatment groups. For $D = 26, 27$ the alternative starts slightly more positive the baseline at childbirth but converges and becomes marginally more negative by $e = 5$. For $D = 29, 30$ the alternative remains consistently more positive than the baseline throughout. For instance, at $D = 30$ and $e = 5$, the baseline equals -0.2 (SE 0.0149) while the alternative equals -0.144 (SE 0.0109), a relative difference of 28%.

Under Assumption NTD-PT, differences between the two series arise from three distinct sources: (i) parallel-trend violations (γ_{PT}), which bias the baseline estimator (Theorem 1); and rescaling by (ii) male normalized effects ($\theta(m, d, a)$) and (iii) counterfactual gender inequality ($\rho(d, \infty, a)$), as shown in (7). The empirical findings, that the alternative estimates are smaller in magnitude compared to the baseline estimates for later treatment groups, are consistent with the predicted direction of rescaling, if males' effects are small and counter-

factual gender inequality is smaller than one.

However, interpreting these patterns as evidence of small bias requires caution. The claim of smaller NTD bias for later treatment groups implies smaller violations of Assumption DID-PT relative to the counterfactual APO. In contrast, Appendix Figure G5 estimates the NTD bias term from Theorem 1 in pre-treatment periods for treatment groups $D \in [26, 30]$. Even for the later treatment groups, the estimated bias terms are non-negligible and display an upward-sloping trend as the age gap between treatment and control widens, suggesting that biases are likely to persist in post-treatment periods.

Next, I consider how imposing an assumption on fathers' effects can be used to identify child penalties via bias-correction, developed in the following subsection.

4.2 Adding a Null Effect Assumption for Fathers

The previous subsection showed that, under the NTD framework, a new causal estimand is identifiable without additional assumptions. This subsection extends the analysis by showing that the original normalized-effects estimand can also be identified when imposing an additional assumption on the magnitude of fathers' earnings effects of parenthood, using a bias-correction formula. The discussion will focus on the specific case of null effects for fathers. Because this assumption cannot be validated in my data, the empirical analysis instead varies the assumed magnitude of fathers' effects to gauge the potential bias in the baseline NTD framework. As before, I first present the theory and then illustrate the bias correction empirically using the Israeli data.

5.2.1. Theory. I begin by outlining the intuition behind how the additional null-effect assumption can be used in both the TD and NTD frameworks to identify the bias and thereby recover the causal estimands of interest for females. I discuss TD first, as the logic may be helpful for understanding the corresponding intuition for NTD. Empirically, however, I evaluate only the NTD case.

When fathers are assumed to be unaffected by childbirth, their observed earnings dynamics capture the parallel-trends violation for men. Since TD assumes that this violation is identical across genders in levels, it can be removed additively from the biased female estimand. Formally, if $ATE(m, d, a) = 0$, then $\delta_{\text{ATE}}(m, d, d', a) = \gamma_{\text{PT}}(m, d, d', a)$. In addition, $\delta_{\text{APO}}(f, d, d', a) = APO(f, d, \infty, a) - \gamma_{\text{PT}}(f, d, d', a)$ (Lemma 1 in Appendix B). Hence, under Assumption TD-PT, and $ATE(m, d, a) = 0$, adding $\delta_{\text{ATE}}(m, d, d', a)$ to $\delta_{\text{APO}}(f, d, d', a)$ identifies $APO(f, d, \infty, a)$.

A similar logic applies for NTD. Assuming fathers are unaffected allows identification of the normalized parallel-trends violation for men. Because NTD assumes that this violation

is identical across genders, it can be removed multiplicatively from the biased estimand for women. Formally, Assumption NTD-PT implies that the ratio between the descriptive δ_{APO} and the true counterfactual APO is the same for men and women. Under the null effect assumption for fathers, this ratio is identified for men, and multiplying the female δ_{APO} by the reciprocal of the male ratio then identifies the female counterfactual APO.

When the bias is identified, a bias-corrected identification formula can be constructed. Theorem 1 shows that under Assumption NTD-PT the descriptive estimand $\delta_\theta(f,d,d',a) - \delta_\theta(m,d,d',a)$ equals the product of a bias parameter, $Bias(d,d',a)$, and the target causal estimand, $\theta(f,d,a) - \theta(m,d,a)$. Under the null-effect assumption for fathers, the bias parameter is identified as

$$Bias(d,d',a) = \frac{APO\left(m,d,\infty,a\right)}{APO\left(m,d,\infty,a\right) - \gamma_{\text{PT}}\left(m,d,d',a\right)} = \frac{\mathbb{E}\left[Y_a \mid G=m, D=d\right]}{\delta_{\text{APO}}\left(m,d,d',a\right)},$$

where the equality of the denominators follows from Lemma 1 in Appendix B, and the equality of the numerators follows from $ATE(m,d,a) = 0$ and consistency. Given this identification of the bias, a bias-corrected identification formula can then be derived, as stated formally in the following result.

Proposition 2. *Assume the same setup as in Theorem 1. If Assumption NTD-PT holds and in addition $ATE(m,d,a) = 0$ then*

$$\theta(f,d,a) = \left(\delta_\theta(f,d,d',a) - \delta_\theta(m,d,d',a)\right) \frac{\delta_{\text{APO}}\left(m,d,d',a\right)}{\mathbb{E}\left[Y_a \mid G=m, D=d\right]}.$$

The proof follows from Theorem 1 and the discussion above on the identification of the bias term. Proposition 2 shows that, under the null-effect assumption, when researchers believe that NTD holds in their context, they can still target the original causal estimand $\theta(f,d,a)$ by multiplying the baseline NTD descriptive estimand, $\delta_\theta(f,d,d',a) - \delta_\theta(m,d,d',a)$, by a bias-correction term that is identifiable from the data.

Whether the additional null assumption holds must be validated. Among the frameworks discussed above, only DID can identify the ATE for fathers. However, Section 3.1 presented theoretical arguments that Assumption DID-PT is likely violated in the child-penalty setting, and Section 3.3 showed that DID fails the validation tests in my application. Results from the broader literature that rely on event-study or DID-based frameworks likewise cannot be used to assess the null assumption because these approaches are subject to the biases discussed in previous sections. Studies that utilize quasi-experimental settings, such as those exploiting the random success of in vitro fertilization (IVF) treatments, may provide more credible evidence, as such randomization mitigates differences between treatment and

control groups. Among these, two studies report results for men: Lundborg et al. (2024) find no significant difference in earnings between successful and unsuccessful IVF treatment couples, while Bensnes et al. (2023) document, if anything, a small positive effect for men.

5.2.2. Empirical Application. As discussed above, the null-effect assumption for fathers cannot be validated. Therefore, I present bias-corrected estimates computed under a range of assumed values for fathers' normalized effect ($\theta(m, d, a)$). Building on (7), I estimate $\widehat{\rho}(d, d, a)$ and $\widehat{\rho}(d, \infty, a)$ as described in Section 4.1, and then compute

$$\frac{(\widehat{\rho}(d, d, a) - \widehat{\rho}(d, \infty, a))}{\widehat{\rho}(d, \infty, a)} (1 + \theta(m, d, a)) . \tag{8}$$

Under NTD and assuming $\theta(m, d, a)$ is known, the estimator in (8) targets $\theta(f, d, a) - \theta(m, d, a)$. Hence, comparing it to $\delta_\theta(f, d, a) - \delta_\theta(m, d, a)$ provides a sense of the magnitude of the bias in the baseline NTD framework.

Appendix Figure G6 presents the results. The black series shows the baseline NTD estimates, as in Figure 4, and the colored series display bias-corrected estimates from Equation (8) for $\theta(m, d, a) \in \{-0.10, -0.05, 0, 0.05, 0.10\}$. The main takeaway is that the baseline NTD estimates lie above the bias-corrected estimates for all selected values of $\theta(m, d, a)$ for treatment groups $D = 24$–27 at later event times. For example, for treatment group $D = 26$, the baseline estimate at five years post-treatment is -0.104 (SE 0.008). Assuming $\theta(m, d, a) = -0.01$ yields a bias-corrected estimate of -0.129, while $\theta(m, d, a) = 0.01$ yields -0.157—that is, 24% and 51% larger in magnitude, respectively. For treatment groups $D \geq 28$, the baseline and bias-corrected estimates are similar.

Connecting this exercise to the theoretical discussion, the difference between the baseline and bias-corrected estimates corresponds to the bias term of the baseline estimator (Theorem 1). The fact that the bias-corrected estimates are more negative for earlier treatment groups, suggesting that bias in the baseline NTD framework attenuates estimates towards zero, is consistent with the discussion in Section 3.1, which argued that parallel-trends violations are likely negative for these groups due to positive selection on human capital.

5 Aggregation

This section discusses aggregating across treatment groups. I first define aggregate causal estimands and then discuss comparisons of aggregated estimates, common in cross-country and subgroup analyses, highlighting how differences in treatment distributions can complicate interpretation. I conclude by illustrating this issue empirically.

5.1 Aggregate Causal Estimands

I begin by defining aggregate causal estimands for normalized effects within gender. For gender $g \in \{f, m\}$, a natural aggregate across treatment groups is

$$\theta_{\text{Agg},1}(g, e) = \mathbb{E}_D \left[\theta(g, D, D + e) \mid G = g, D + e < D_{\max}\right].$$

The aggregated estimand $\theta_{\text{Agg},1}(g, e)$ averages normalized effects for individuals of gender g at event time e, across all treatment groups which are observed e periods post childbirth.[18] However, the event study estimator $\widehat{\theta}_{\text{ES}}(g, e)$ in (2) is a ratio of averages, whereas $\theta_{\text{Agg},1}(g, e)$ is an average of ratios. An aggregate estimand that matches the structure of $\widehat{\theta}_{\text{ES}}(g, e)$ is

$$\theta_{\text{Agg},2}(g, e) = \frac{\mathbb{E}_D \left[ATE(g, D, D + e) \mid G = g, D + e < D_{\max}\right]}{\mathbb{E}_D \left[APO(g, D, \infty, D + e) \mid G = g, D + e < D_{\max}\right]}.$$

Comparing $\theta_{\text{Agg},1}(g, e)$ to $\theta_{\text{Agg},2}(g, e)$, while the first uses the treatment distribution to summarize effects, the latter effectively gives higher weight to treatment groups with higher counterfactual APOs, which are typically the later-treated, as their post-treatment periods occur in later parts of the life-cycle compared to earlier treatment groups.[19] Since giving higher weights to higher earning treatment groups at time of treatment does not seem warranted, I see $\theta_{\text{Agg},1}(g, e)$ as the preferable aggregate estimand.

Finally, an aggregate version of the gender-inequality estimand from Section 4.1 can be written as

$$\rho_{\text{Agg}}(e) = \mathbb{E}_D \big[\Delta\rho(D, D + e) \mid D + e < D_{\max}\big].$$

5.2 Comparing Aggregates

Having defined potential target aggregate estimands, I next discuss issues that arise when aggregating estimates, and when comparing aggregates across groups.

First, as discussed above, under either DID or NTD the single-treatment-group estimates incorporates both effect and bias. Hence aggregating single-treatment-group estimates across

[18] Comparing values across e combines variation in causal effects with shifts in treatment group composition, as noted by Callaway and Sant'Anna (2021). This can be addressed by conditioning on a maximum exposure length e', and comparing $e_1, e_2 \leq e'$.

[19] Let $p_d = \Pr(D = d \mid G = g,\ D + e < D_{\max})$, $\text{ATE}_d = \text{ATE}(g, d, D + e)$, $\text{APO}_d = \text{APO}(g, d, \infty, D + e)$, and $\theta_d = \text{ATE}_d / \text{APO}_d$. Then

$$\theta_{\text{Agg},2}(g, e) = \frac{\sum_d p_d \,\text{ATE}_d}{\sum_d p_d \,\text{APO}_d} = \frac{\sum_d p_d \,\text{APO}_d \,\theta_d}{\sum_d p_d \,\text{APO}_d} = \sum_d w_d \,\theta_d, \quad w_d = \frac{p_d \,\text{APO}_d}{\sum_{d'} p_{d'} \,\text{APO}_{d'}}.$$

Hence $\theta_{\text{Agg},2}$ gives higher weights to groups with higher APOs.

multiple treatment groups combines heterogeneity in both effects and biases, implying that differences in aggregates may arise from variation in biases across treatment groups. This statement holds for normalized-event-study estimates as well, $\widehat{\theta}_{\text{ES}}(g, e)$ in (2), and is a distinct source of bias, arising from aggregation of biases due to parallel trend violations, rather than the estimation bias due to variation in treatment time discussed in Section 1.1.

Second, even abstracting from bias, comparisons of aggregates across different strata, such as countries or parent types, are difficult to interpret as observed differences may reflect differences in either single-treatment-group effects or treatment distributions.[20] Hence, even if effects for treatment group d are identical across strata but heterogeneous across treatment groups, countries may still differ in their aggregates due to differences in treatment distributions.

Two exercises may help interpreting differences in aggregates. First, compare disaggregated estimates across countries. Second, construct aggregated estimates using a fixed reference distribution held constant across countries. Such exercises may shed light on the mechanisms underlying variation in aggregate estimates across strata.

5.3 Empirical Illustration

I conclude this section by illustrating how treatment distributions might differ across strata, and show that even if single-treatment-group effects are constant across countries, differences in treatment distributions may affect the aggregated results and their interpretation.

Appendix Figure G7 documents cross-country variation in the distribution of age at first childbirth for five OECD countries.: the United States exhibits a right-skewed distribution of first-birth ages; Denmark and Sweden are more centered; Italy and Spain are left-skewed. Accordingly, the United States may display a different aggregate than Italy simply because it places greater weight on earlier treatment groups.

To illustrate how differences in treatment distributions may affect results, Appendix Figure G8 reports aggregated estimates of the effect of parenthood on the gender earnings ratio, i.e., the estimated $\rho_{\text{Agg}}(e)$ defined in Section 5.1, using three different treatment distributions. These distributions are motivated from the previous discussion on OECD distributions: one gives more weight to earlier treatment groups (right-skewed), the second gives more weight to mid-range treatment groups (centered), and the third places more weight on later treatment groups (left-skewed).

[20] As examples of comparisons of aggregates, Kleven, Landais, Posch, et al. (2019) compare between countries, Kleven et al. (2021) compare between biological and adoptive parents, Andresen and Nix (2022) compare between heterosexual and lesbian couples, and and Jensen et al. (2024) compare between low- and high-educated individuals. Each of these calculates normalized event studies within each strata, i.e., country or parent type, and then compares (aggregated) estimates across strata.

The documented aggregated estimates show two different patterns. Under the right-skewed distribution, the effect of parenthood on gender inequality in earnings becomes smaller over time almost linearly. Fitting a line to these estimates, one can expect the effect to be zero at event time 10. In contrast, under the left-skewed distribution the aggregate drops from 27% to 16% between event times zero and two, but then stays around 15% up to event time five. That is, under this distribution, one can conjecture that the effect of parenthood on gender inequality does not recede over time. This exercise illustrates that when single-treatment-group effects are heterogeneous, differences in treatment distributions can substantially affect aggregated estimators.

6 Conclusion

This paper revisits the identification of child penalties in the normalized event-study framework. I show that the implicit identifying assumption, gender equality in normalized parallel-trend violations, does not identify the target causal estimand, the gender gap in normalized effects. The analysis then proceeds in two directions. The first explores alternative frameworks, namely difference-in-differences and triple differences, and examines potential violations arising from selection into treatment and life-cycle variation in counterfactual gender inequality. The second aims to recover identification within the normalized framework. I show that the effect of parenthood on the gender earnings ratio is identified without further assumptions, suggesting it as a potential target estimand for future research. I also discuss how assumptions about fathers' effects can be used to identify mothers' normalized effects through a bias-correction approach. The paper concludes by examining aggregation across multiple treatment.

While this paper provides a better understanding of the identification of child penalties, several avenues for future research remain. The analysis relies on both SUTVA and Assumption NA; exploring how violations of these assumptions affect results is an important next step. Moreover, the discussion focused on earnings as the outcome; extending the framework to employment, hours worked, or firm-level dynamics offers a natural direction for future work. More broadly, the approach developed here can inform other applications with endogenous treatment timing, where selection on treatment and counterfactual inequality affect the validity of identification assumptions.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, *88*(1), 265–296.

Abbring, J. H., & Van den Berg, G. J. (2003). The nonparametric identification of treatment effects in duration models. *Econometrica*, *71*(5), 1491–1517.

Adda, J., Dustmann, C., & Stevens, K. (2017). The career costs of children. *Journal of Political Economy*, *125*(2), 293–337.

Andresen, M. E., & Nix, E. (2022). What causes the child penalty? evidence from adopting and same-sex couples. *Journal of labor economics*, *40*(4), 971–1004.

Angelov, N., Johansson, P., & Lindahl, E. (2016). Parenthood and the gender gap in pay. *Journal of labor economics*, *34*(3), 545–579.

Angrist, J., & Evans, W. N. (1996). Children and their parents' labor supply: Evidence from exogenous variation in family size.

Angrist, J., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Attanasio, O., Low, H., & Sánchez-Marcos, V. (2008). Explaining changes in female labor supply in a life-cycle model. *American Economic Review*, *98*(4), 1517–1552.

Autor, D. H. (2003). Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing. *Journal of labor economics*, *21*(1), 1–42.

Bearth, N. (2024). Beyond baby blues: The child penalty in mental health in switzerland. *arXiv preprint arXiv:2410.20861*.

Becker, G. S., Murphy, K. M., & Tamura, R. (1990). Human capital, fertility, and economic growth. *Journal of political economy*, *98*(5, Part 2), S12–S37.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *Journal of political economy*, *75*(4, Part 1), 352–365.

Bensnes, S., Huitfeldt, I., & Leuven, E. (2023). *Reconciling estimates of the long-term earnings effect of fertility* (tech. rep.). Discussion Papers.

Bertrand, M., Goldin, C., & Katz, L. F. (2010). Dynamics of the gender gap for young professionals in the financial and corporate sectors. *American economic journal: applied economics*, *2*(3), 228–255.

Black, S. E., Devereux, P. J., & Salvanes, K. G. (2008). Staying in the classroom and out of the maternity ward? the effect of compulsory schooling laws on teenage births. *The economic journal*, *118*(530), 1025–1054.

Blackburn, M. L., Bloom, D. E., & Neumark, D. (1993). Fertility timing, wages, and human capital. *Journal of Population Economics*, *6*(1), 1–30.

Blundell, R., Pistaferri, L., & Saporta-Eksten, I. (2018). Children, time allocation, and consumption insurance. *Journal of Political Economy*, *126*(S1), S73–S115.

Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies*, rdae007.

Buckles, K. (2008). Understanding the returns to delayed childbearing for working women. *American Economic Review*, *98*(2), 403–407.

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of econometrics*, *225*(2), 200–230.

Cengiz, D., Dube, A., Lindner, A., & Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics*, *134*(3), 1405–1454.

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The quarterly journal of economics*, *129*(4), 1553–1623.

Cortés, P., & Pan, J. (2023). Children and the remaining gender gaps in the labor market. *Journal of Economic Literature*, *61*(4), 1359–1409.

Cristia, J. P. (2008). The effect of a first child on female labor supply: Evidence from women seeking fertility services. *Journal of Human Resources*, *43*(3), 487–510.

Cunha, F., & Heckman, J. (2007). The technology of skill formation. *American economic review*, *97*(2), 31–47.

De Chaisemartin, C., & d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964–2996.

De La Croix, D., & Doepke, M. (2003). Inequality and growth: Why differential fertility matters. *American Economic Review*, *93*(4), 1091–1113.

de la Vega, N. (2022). The differential effect of childbirth on men's and women's careers. *Labour Economics*, *78*, 102249.

Doepke, M., & Kindermann, F. (2019). Bargaining over babies: Theory, evidence, and policy implications. *American Economic Review*, *109*(9), 3264–3306.

Eckstein, Z., Keane, M., & Lifshitz, O. (2019). Career and family decisions: Cohorts born 1935–1975. *Econometrica*, *87*(1), 217–253.

Fajardo-Gonzalez, J., Hasanbasri, A., & Rios-Avila, F. (2024). *Is there a gendered parenthood penalty in indonesian labor markets?* World Bank.

Francesconi, M. (2002). A joint dynamic model of fertility and work of married women. *Journal of labor Economics*, *20*(2), 336–380.

Geronimus, A. T., & Korenman, S. (1992). The socioeconomic consequences of teen childbearing reconsidered. *The Quarterly Journal of Economics*, *107*(4), 1187–1214.

Goldin, C. (2024). Nobel lecture: An evolving economic force. *American Economic Review*, *114*(6), 1515–1539.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277.

Gould, E. D., & Lichtinger, G. (2024, November). *Child penalties, child outcomes, and family culture* (IZA Discussion Paper No. 17455) (IZA Discussion Paper No. 17455). IZA – Institute of Labor Economics. Bonn, Germany.

Gronau, R. (1977). Leisure, home production, and work–the theory of the allocation of time revisited. *Journal of political economy*, *85*(6), 1099–1123.

Hazan, M., & Zoabi, H. (2015). Do highly educated women choose smaller families? *The Economic Journal*, *125*(587), 1191–1226.

Heckman, J. J. (1976). A life-cycle model of earnings, learning, and consumption. *Journal of political economy*, *84*(4, Part 2), S9–S44.

Heckman, J. J., & Mosso, S. (2014). The economics of human development and social mobility. *Annu. Rev. Econ.*, *6*(1), 689–733.

Jakobsen, K. M., Jørgensen, T. H., & Low, H. (2022). Fertility and family labor supply.

Jensen, M., Adams, A., & Petrongolo, B. (2024). *Birth timing and spacing: Implications for parental leave dynamics and child penalties* (tech. rep.).

Keane, M. P., & Wolpin, K. I. (2010). The role of labor and marriage markets, preference heterogeneity, and the welfare system in the life cycle decisions of black, hispanic, and white women. *International Economic Review*, *51*(3), 851–892.

Kleven, H. (2022). *The geography of child penalties and gender norms: Evidence from the united states* (tech. rep.). National Bureau of Economic Research.

Kleven, H., Landais, C., Posch, J., Steinhauer, A., & Zweimuller, J. (2019). Child penalties across countries: Evidence and explanations. *AEA Papers and Proceedings*, *109*, 122–126.

Kleven, H., Landais, C., & Sogaard, J. (2019). Children and gender inequality: Evidence from denmark. *American Economic Journal: Applied Economics*, *11*(4), 181–209.

Kleven, H., Landais, C., & Søgaard, J. E. (2021). Does biology drive child penalties? evidence from biological and adoptive families. *American Economic Review: Insights*, *3*(2), 183–198.

Leventer, D. (2025). Conditional triple difference-in-differences. *arXiv preprint arXiv:2502.16126*.

Lin, X. (2025). Long-term child penalties and mothers' age at first birth.

Lundborg, P., Plug, E., & Rasmussen, A. W. (2024). *Is there really a child penalty in the long run? new evidence from ivf treatments* (tech. rep.). IZA Discussion Papers.

McCrary, J., & Royer, H. (2011). The effect of female education on fertility and infant health: Evidence from school entry policies using exact date of birth. *American economic review*, *101*(1), 158–195.

Melentyeva, V., & Riedel, L. (2023). *Child penalty estimation and mothers' age at first birth* (tech. rep.). ECONtribute Discussion Paper.

Moffitt, R. (1984). Profiles of fertility, labour supply and wages of married women: A complete life-cycle model. *The Review of Economic Studies*, *51*(2), 263–278.

Msemburi, W., Karlinsky, A., Knutson, V., Aleshin-Guendel, S., Chatterji, S., & Wakefield, J. (2023). The who estimates of excess mortality associated with the covid-19 pandemic. *Nature*, *613*(7942), 130–137.

Orenstein, W. A., Bernier, R. H., Dondero, T. J., Hinman, A. R., Marks, J. S., Bart, K. J., & Sirotkin, B. (1985). Field evaluation of vaccine efficacy. *Bulletin of the World Health Organization*, *63*(6), 1055.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, *4*(3), 305–322.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, *75*(371), 591–593.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175–199.

VanderWeele, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology*, *20*(6), 880–883.

Wing, C., Freedman, S. M., & Hollingsworth, A. (2024). *Stacked difference-in-differences* (tech. rep.). National Bureau of Economic Research.

Yakin, E. (2021). *Parenthood and gender identity impacts on women labor force outcomes in israel* (tech. rep.). The Hebrew University of Jerusalem.

A Appendix Data

A.1 Variable Definitions

Grandparents earnings rank. For each parent in the data, I link the identifiers of her biological father and mother, denoted as the grandfather and grandmother. For each parent, I sum the grandparents' total annual earnings over the years in which the parent was aged 5–10 and divide this sum by six to obtain mean household earnings during that period. I then rank mean household earnings within the parent's birth cohort to construct the grandparents' earnings-rank variable. This procedure follows standard practice in the intergenerational

income-mobility literature (e.g., Chetty et al., 2014).

Years of education and highest degree. The CBS maintains an annually updated dataset recording individuals' years of schooling. We use this to construct a variable for the maximum observed years of education for each individual. We also construct a binary indicator for whether the individual holds a bachelor's degree or higher.

Meitsav test score. The Meitzav is a national standardized exam administered by the Israeli Ministry of Education in four subjects: science, mathematics, English, and Hebrew, first implemented in 2002. For each individual with available data, I retain the mathematics score and standardize it to have mean zero and standard deviation one within each exam year.

High-school credits. In Israel, high-school students complete subjects at varying levels, defined by the number of credit units in each subject up to a maximum of five. For each individual, I construct binary indicators for completing five credit units in selected subjects: mathematics, English, computers, and physics, as reported in the individual's matriculation certificate.

UPET score. Admission to Israeli universities and colleges is primarily based on performance in the University Psychometric Entrance Test (UPET). For each individual, we construct a variable recording their maximum score, and a binary indicator for whether they ever took the test. While the UPET score is often used as a measure of ability, it should be interpreted cautiously, as the test is commonly taken at older ages and after paid preparation courses. And, students have varying target test scores based on varying thresholds of university fields of study.

Ethnicity and religion. The Civil Registry classifies individuals as Jewish or Arab. Among Jewish individuals, religious affiliation is inferred (with some measurement error) from school type, which falls into one of three streams: Ultra-Orthodox (Haredi), state-religious, and state-secular.

A.2 Analysis Dataset Definition

The main sample restrictions are described in Section 1.2. Two additional exclusions, omitted from the main text due to their negligible impact on sample size, are applied throughout the analysis. First, I drop cases in which the recorded birth year of the parent is later than that of their first child (951 observations). Second, I drop individuals recorded as giving birth at age ten or younger (443 observations).

The definitions below describe the construction of the auxiliary datasets used for Figure 1 in Section 3.1. Because these figures do not rely on observed parental earnings, the treatment-

group range is extended to all first-birth ages between 20 and 40 (rather than 24–40 in the main analysis).

I now turn to discussing the dataset definitions that construct Figure 1 in Section 3.1. Since these figures do not rely on the earnings of the parents, I show results for a wider range of treatment groups, namely 20-40, differently from the analysis sample, which focuses on 24-40.

Grandparents earnings rank. Grandparents' earnings come from administrative income data for 1990–2020. Because earnings data begin in 1990, the youngest parent birth cohort for which the grandparents' earnings rank can be constructed is 1985. Since the parent sample extends up to the 1990 birth cohort, the dataset used for this variable includes cohorts 1985–1990. Given that income data end in 2020, the latest observable first-birth age (treatment group) for which grandparents' earnings rank is available is 35.

Grandparents Education. For each grandparent, I take the reported years of education and classify it into three categories: missing, at most 12 years (high school or less), or above 12 years (post-secondary education). This variable is available for all treatment groups with first-birth ages between 20 and 40. The share of missing observations ranges between 1–2% for grandmothers and 3–5% for grandfathers across treatment groups.

Meitsav. Data on Meitzav test scores are available for the years 2002–2019. The exams are administered in 5th and 8th grades, corresponding to ages 10–11 and 13–14, respectively. The 5th-grade test is observable for birth cohorts 1991 and onward. To analyze 5th-grade scores, I therefore include two additional cohorts not used in the main analysis sample: 1991 and 1992. Given that birth data is available up to 2020, this measure is available for treatment groups with first-birth ages 20–29. The 8th-grade test is observable for birth cohorts 1988 and onward. To analyze 8th-grade scores, I limit the dataset to birth cohorts 1988–1990, and hence the measure is available for treatment groups aged 20–32.

High-school credits. Data on the number of credit units in high-school subjects are available for most of the school cohort beginning with the 1980 birth cohort. Accordingly, this measure is constructed for all treatment groups with first-birth ages between 20 and 40. Note that sample size declines sharply at higher first-birth ages: the number of observations decreases from 25,710 for treatment age 30, to 4,451 for treatment age 35, and to 248 for treatment age 40.

B Appendix Identification

Before stating the main result, I introduce a lemma that links the descriptive estimands to their causal counterparts and to potential differences in counterfactual trends between the

treatment and control groups.

Lemma 1. *Assume the setup in Theorem 1. If Assumption NA holds for both $g \in \{f, m\}$ and treatment groups $\{d, d'\}$, then*

$$\begin{aligned}
\delta_{\text{APO}}(g, d, d', a) &= APO(g, d, \infty, a) - \gamma_{\text{PT}}(g, d, d', a), \\
\delta_{\text{ATE}}(g, d, d', a) &= ATE(g, d, a) + \gamma_{\text{PT}}(g, d, d', a), \\
\delta_{\theta}(g, d, d', a) &= \theta(g, d, a) \frac{APO(g, d, \infty, a)}{APO(g, d, \infty, a) - \gamma_{\text{PT}}(g, d, d', a)} \\
&\quad + \frac{\gamma_{\text{PT}}(g, d, d', a)}{APO(g, d, \infty, a) - \gamma_{\text{PT}}(g, d, d', a)}.
\end{aligned}$$

Proof of Lemma 1. Re-arranging the definition of γ_{PT} implies

$$\begin{aligned}
\mathbb{E}\left[Y_a\left(\infty\right) \mid G = g, D = d\right] = \gamma_{\text{PT}}\left(g, d, d', a\right) &+ \mathbb{E}\left[Y_{d-1}\left(\infty\right) \mid G = g, D = d\right] \\
&+ \mathbb{E}\left[Y_{i,a}\left(\infty\right) - Y_{d-1}\left(\infty\right) \mid G = g, D = d'\right].
\end{aligned}$$

Assumption NA and consistency imply

$$\begin{aligned}
\mathbb{E}\left[Y_a\left(\infty\right) \mid G = g, D = d\right] = \gamma_{\text{PT}}\left(g, d, d', a\right) &+ \mathbb{E}\left[Y_{d-1} \mid G = g, D = d\right] \\
&+ \mathbb{E}\left[Y_a - Y_{d-1} \mid G = g, D = d'\right].
\end{aligned}$$

Substituting the definition of δ_{APO} and re-arranging we obtain

$$\delta_{\text{APO}}\left(g, d, d', a\right) = APO\left(g, d, \infty, a\right) - \gamma_{\text{PT}}\left(g, d, d', a\right) \tag{B1}$$

Substituting into the definition of the δ_{ATE}

$$\begin{aligned}
\delta_{\text{ATE}}\left(g, d, d', a\right) &= \mathbb{E}\left[Y_{i,a} \mid G_i = g, D_i = d\right] - \left(APO\left(g, d, \infty, a\right) - \gamma_{\text{PT}}\left(g, d, d', a\right)\right) \\
&= ATE\left(g, d, a\right) + \gamma_{\text{PT}}\left(g, d, d', a\right).
\end{aligned} \tag{B2}$$

Finally, substitute (B1) and (B2) into the definition of δ_θ to obtain

$$\begin{aligned}
\delta_\theta(g,d,d',a) &= \frac{\delta_{\text{ATE}}(g,d,d',a)}{\delta_{\text{APO}}(g,d,d',a)} \\
&= \frac{ATE(g,d,a)+\gamma_{\text{PT}}(g,d,d',a)}{APO(g,d,\infty,a)-\gamma_{\text{PT}}(g,d,d',a)} \\
&= \frac{\theta(g,d,a)\,APO(g,d,\infty,a)+\gamma_{\text{PT}}(g,d,d',a)}{APO(g,d,\infty,a)-\gamma_{\text{PT}}(g,d,d',a)} \\
&= \theta(g,d,a)\frac{APO(g,d,\infty,a)}{APO(g,d,\infty,a)-\gamma_{\text{PT}}(g,d,d',a)} \\
&+ \frac{\gamma_{\text{PT}}(g,d,d',a)}{APO(g,d,\infty,a)-\gamma_{\text{PT}}(g,d,d',a)}. \qquad \text{(B3)}
\end{aligned}$$

□

Proof of Proposition 1. Considering $\delta_\theta(f,d,d',a)=\delta_\theta(m,d,d',a)$ for $a<d,d'$ we obtain

$$\begin{aligned}
&\frac{\delta_{\text{ATE}}(f,d,d',a)}{\delta_{\text{APO}}(f,d,d',a)} = \frac{\delta_{\text{ATE}}(m,d,d',a)}{\delta_{\text{APO}}(m,d,d',a)} \\
\leftrightarrow &\frac{ATE(f,d,a)+\gamma_{\text{PT}}(f,d,d',a)}{APO(F,d,\infty,a)-\gamma_{\text{PT}}(F,d,d',a)} = \frac{ATE(m,d,a)+\gamma_{\text{PT}}(m,d,d',a)}{APO(m,d,\infty,a)-\gamma_{\text{PT}}(m,d,d',a)} \\
\leftrightarrow &\frac{\gamma_{\text{PT}}(f,d,d',a)}{APO(f,d,\infty,a)-\gamma_{\text{PT}}(f,d,d',a)} = \frac{\gamma_{\text{PT}}(m,d,d',a)}{APO(m,d,\infty,a)-\gamma_{\text{PT}}(m,d,d',a)} \\
\leftrightarrow &\gamma_{\text{PT}}(f,d,d',a)\left[APO(m,d,\infty,a)-\gamma_{\text{PT}}(m,d,d',a)\right] \\
&= \gamma_{\text{PT}}(m,d,d',a)\left[APO(f,d,\infty,a)-\gamma_{\text{PT}}(f,d,d',a)\right] \\
\leftrightarrow &\gamma_{\text{PT}}(f,d,d',a)\,APO(m,d,\infty,a) \\
&= \gamma_{\text{PT}}(m,d,d',a)\,APO(f,d,\infty,a) \\
\leftrightarrow &\frac{\gamma_{\text{PT}}(f,d,d',a)}{APO(f,d,\infty,a)} = \frac{\gamma_{\text{PT}}(m,d,d',a)}{APO(m,d,\infty,a)},
\end{aligned}$$

where the first iff is due to substituting (B1) and (B2), the second iff follows from Assumption NA, and the rest result from algebra. □

Proof of Theorem 1. Assumption NTD-PT implies

$$\begin{aligned}\frac{APO\left(f,d,\infty,a\right)}{APO\left(f,d,\infty,a\right)-\gamma_{\mathrm{PT}}\left(f,d,d',a\right)} &= \frac{APO\left(f,d,\infty,a\right)}{APO\left(f,d,\infty,a\right)-\frac{APO(f,d,\infty,a)}{APO(m,d,\infty,a)}\gamma_{\mathrm{PT}}\left(m,d,d',a\right)} \\ &= \frac{1}{1-\frac{\gamma_{\mathrm{PT}}(m,d,d',a)}{APO(m,d,\infty,a)}} \\ &= \frac{APO\left(m,d,\infty,a\right)}{APO\left(m,d,\infty,a\right)-\gamma_{\mathrm{PT}}\left(m,d,d',a\right)}, \end{aligned} \tag{B4}$$

and

$$\begin{aligned}\frac{\gamma_{\mathrm{PT}}\left(f,d,d',a\right)}{APO\left(f,d,\infty,a\right)-\gamma_{\mathrm{PT}}\left(f,d,d',a\right)} &= \frac{\frac{APO(f,d,\infty,a)}{APO(m,d,\infty,a)}\gamma_{\mathrm{PT}}\left(m,d,d',a\right)}{APO\left(f,d,\infty,a\right)-\frac{APO(f,d,\infty,a)}{APO(m,d,\infty,a)}\gamma_{\mathrm{PT}}\left(m,d,d',a\right)} \\ &= \frac{\frac{\gamma_{\mathrm{PT}}(m,d,d',a)}{APO(m,d,\infty,a)}}{1-\frac{\gamma_{\mathrm{PT}}(m,d,d',a)}{APO(m,d,\infty,a)}} \\ &= \frac{\gamma_{\mathrm{PT}}\left(m,d,d',a\right)}{APO\left(m,d,\infty,a\right)-\gamma_{\mathrm{PT}}\left(m,d,d',a\right)}. \end{aligned} \tag{B5}$$

This shows that $Bias(d,d',a)$, as stated in the theorem, is indeed gender invariant. Subtracting (B3) using $g=f$ from (B3) using $g=m$, and applying (B4) and (B5) we obtain the result in the theorem, which concludes the proof. □

The next lemma characterizes the bias for another target causal estimand that is sometimes used in the literature to quantify child penalties.

Lemma 2. *Assume the same setup as in Theorem 1. Let*

$$\begin{aligned}P(d,a) &= \frac{ATE(f,d,a)-ATE(m,d,a)}{APO(f,d,\infty,a)}, \\ \delta_P(d,d',a) &= \frac{\delta_{\mathrm{ATE}}(f,d,d',a)-\delta_{\mathrm{ATE}}(m,d,d',a)}{\delta_{\mathrm{APO}}(f,d,d',a)}.\end{aligned}$$

Then,

$$\delta_P(d,d',a) = P(d,a)\times Bias_1(d,d',a)+Bias_2(d,d',a).$$

where $Bias_1(d,d',a)$ is equal to $Bias(d,d',a)$ from Theorem 1, and

$$Bias_2(d,d',a) = \frac{\gamma_{\mathrm{PT}}(f,d,d',a)-\gamma_{\mathrm{PT}}(m,d,d',a)}{APO(f,d,\infty,a)-\gamma_{\mathrm{PT}}(f,d,d',a)}.$$

Proof. From Lemma 1

$$\delta_P(d,d',a) = \frac{ATE(f,d,a) + \gamma_{\text{PT}}(f,d,d',a)}{APO(f,d,\infty,a) - \gamma_{\text{PT}}(f,d,d',a)} - \frac{ATE(m,d,a) + \gamma_{\text{PT}}(m,d,d',a)}{APO(f,d,\infty,a) - \gamma_{\text{PT}}(f,d,d',a)}. \tag{B6}$$

Adding and subtracting $P(d,a)$ in (B6), after some algebra, yields

$$\begin{aligned}\delta_P(d,d',a) = P(d,a) &+ \frac{\gamma_{\text{PT}}(f,d,d',a)}{APO(f,d,\infty,a)} \frac{ATE(f,d,a) - ATE(m,d,a)}{APO(f,d,\infty,a) - \gamma_{\text{PT}}(f,d,d',a)} \\ &+ \frac{\gamma_{\text{PT}}(f,d,d',a) - \gamma_{\text{PT}}(m,d,d',a)}{APO(f,d,\infty,a) - \gamma_{\text{PT}}(f,d,d',a)}.\end{aligned}$$

Rearranging terms we obtain

$$\begin{aligned}\delta_P(d,d',a) = P(d,a) \times &\frac{APO(f,d,\infty,a)}{APO(f,d,\infty,a) - \gamma_{\text{PT}}(f,d,d',a)} \\ &+ \frac{\gamma_{\text{PT}}(f,d,d',a) - \gamma_{\text{PT}}(m,d,d',a)}{APO(f,d,\infty,a) - \gamma_{\text{PT}}(f,d,d',a)},\end{aligned}$$

which concludes the proof. □

Proof of Theorem 2. Substituting definitions, Assumption NTD-PT can be written as

$$\begin{aligned}&\frac{APO\left(f,d,\infty,a\right) - APO\left(f,d,\infty,d-1\right) - \left[APO\left(f,d',\infty,a\right) - APO\left(f,d',\infty,d-1\right)\right]}{APO\left(f,d,\infty,a\right)} \\ &= \frac{APO\left(m,d,\infty,a\right) - APO\left(m,d,\infty,d-1\right) - \left[APO\left(m,d',\infty,a\right) - APO\left(m,d',\infty,d-1\right)\right]}{APO\left(m,d,\infty,a\right)},\end{aligned}$$

which can be simplified into

$$\begin{aligned}&\frac{APO\left(f,d,\infty,d-1\right) + \left[APO\left(f,d',\infty,a\right) - APO\left(f,d',\infty,d-1\right)\right]}{APO\left(f,d,\infty,a\right)} \\ &= \frac{APO\left(m,d,\infty,d-1\right) + \left[APO\left(m,d',\infty,a\right) - APO\left(m,d',\infty,d-1\right)\right]}{APO\left(m,d,\infty,a\right)}.\end{aligned} \tag{B7}$$

Substituting definitions into (B7) this

$$\frac{\delta_{\text{APO}}(f,d,d',a)}{\delta_{\text{APO}}(m,d,d',a)} = \frac{APO\left(f,d,\infty,a\right)}{APO\left(m,d,\infty,a\right)},$$

and consistency implies

$$\frac{APO\left(f,d,d,a\right)}{APO\left(m,d,d,a\right)}=\frac{\mathbb{E}\left[Y_a\mid G=f,D=d\right]}{\mathbb{E}\left[Y_a\mid G=m,D=d\right]}.$$

This completes the proof of the theorem. I furthermore show how to derive the connection to the original NTD causal estimand.

$$\begin{aligned}
&\frac{APO\left(f,d,d,a\right)}{APO\left(m,d,d,a\right)}-\frac{APO\left(f,d,\infty,a\right)}{APO\left(m,d,\infty,a\right)}\\
&=\frac{ATE(f,d,a)+APO\left(f,d,\infty,a\right)}{ATE(m,d,a)+APO\left(m,d,\infty,a\right)}-\frac{APO\left(f,d,\infty,a\right)}{APO\left(m,d,\infty,a\right)}\\
&=\frac{\theta(f,d,a)APO\left(f,d,\infty,a\right)+APO\left(f,d,\infty,a\right)}{\theta(m,d,a)APO\left(f,d,\infty,a\right)+APO\left(m,d,\infty,a\right)}-\frac{APO\left(f,d,\infty,a\right)}{APO\left(m,d,\infty,a\right)}\\
&=\frac{APO\left(f,d,\infty,a\right)\left[\theta(f,d,a)+1\right]}{APO\left(m,d,\infty,a\right)\left[\theta(m,d,a)+1\right]}-\frac{APO\left(f,d,\infty,a\right)}{APO\left(m,d,\infty,a\right)}\\
&=\frac{APO\left(f,d,\infty,a\right)}{APO\left(m,d,\infty,a\right)}\left[\frac{\left[\theta(f,d,a)+1\right]}{\left[\theta(m,d,a)+1\right]}-1\right]\\
&=\frac{APO\left(f,d,\infty,a\right)}{APO\left(m,d,\infty,a\right)}\left[\frac{\theta\left(f,d,a\right)-\theta\left(m,d,a\right)}{\theta\left(m,d,a\right)+1}\right]
\end{aligned}$$

□

Proof of Proposition 2. From $ATE(m,d,a)=0$ we obtain $APO(m,d,\infty,a)=APO(m,d,d,a)$. From consistency $\mathbb{E}\left[Y_a\mid G=m,D=d\right]=APO(m,d,d,a)$. Hence $\mathbb{E}\left[Y_a\mid G=m,D=d\right]=APO(m,d,\infty,a)$. Combined with (B1) this implies

$$\frac{\mathbb{E}\left[Y_a\mid G=m,D=d\right]}{\delta_{\text{APO}}(m,d,d',a)}=\frac{APO(m,d,\infty,a)}{APO(m,d,\infty,a)-\gamma_{\text{PT}}\left(m,d,d',a\right)}. \tag{B8}$$

This shows that $\frac{\mathbb{E}[Y_a|G=m,D=d]}{\delta_{\text{APO}}(m,d,d',a)}$ is equal to $Bias(d,d',a)$ from Theorem 1. Hence combining Theorem 1 and (B8) imply that dividing $\delta_\theta(f,d,d',a)-\delta_\theta(m,d,d',a)$ by $\frac{\mathbb{E}[Y_a|G=m,D=d]}{\delta_{\text{APO}}(m,d,d',a)}$ identifies $\theta(f,d,a)$, noting that $\theta(m,d,a)=0$ from $ATE(m,d,a)=0$. □

C Appendix TD

In this appendix I present an empirical exercise to suggest at the bias in TD due to violations of the identification assumption. The exercise proceeds in two steps for 2×2 with treatment group d and target age a:

1. Use pre-childbirth data and estimate mean earnings by gender, treatment group, and

age using sample analogs, and substitute these values into equation (6).

2. Estimate $\widehat{\rho}(d, d-1)$. Next, estimate the change $\rho(d', d-1) - \rho(d', a)$ for later treatment groups ($d' > a$), average these changes, and add them to $\widehat{\rho}(d, d-1)$ to impute $\rho(d, a)$. This step is motivated by the above empirical evidence, which shows a common life-cycle trajectory for ρ.

After these steps, only one unknown remains in equation (6), namely $APO(m, d, \infty, a)$. This allows to solve for the value of $APO(m, d, \infty, a)$ that satisfies the TD identifying assumption, denoted $APO^{\star}$. In the Israeli application below, I compute this value and use it to provide a suggestive indication of the sign of TD bias. The same procedure can also be applied by inputting plausible values of $APO(m, d, \infty, a)$, which delivers bounds on the bias resulting from violations of the assumption.

Appendix Table G1 reports $APO^{\star}$, the imputed $\rho(d, \infty, a)$, and the other estimated components of the exercise for early treatment groups $D \in \{24, 25, 26\}$ and event times $e \in \{3, 4, 5\}$. $\rho(d, \infty, a)$ is imputed using later treatment groups 29-35 only when those are pre-treatment at a. Given the signs of Terms 1–4, the sum is positive if the true value of $APO(m, d, \infty, a)$ is below $APO^{\star}$, and negative if it exceeds it. Strikingly, across all cases the threshold $APO^{\star}$ exceeds the observed APO for the control group. Since the control group is assumed to be positively selected relative to the treatment group, it is likely that $APO(m, d, \infty, a) < APO(m, d', \infty, a)$, and hence that $APO(m, d, \infty, a) < APO^{\star}$. This suggests that the sum on the right-hand side of (6) is positive, so TD exceeds the true difference in average treatment effects (Lemma 1 Appendix B). In other words, in the considered application, TD likely understates the gender gap in the effect of parenthood on earnings in levels, among early treatment groups at later ages. Because this result relies on data, it may differ across settings. But the exercise itself is easily replicable, providing a general strategy for assessing TD violations in child penalty applications.

D Appendix Estimation and Inference

For a triplet of treatment group, control group, and target age (d, d', a), the main text considers fifteen distinct descriptive estimands excluding methods that condition on covariates. Specifically, DID includes three for each gender (six in total), TD and NTD contribute one each, the NTD alternative identifies an additional normalized estimand, and—under the null–effect–for–fathers assumption—TD and NTD also identify the causal estimands APO, ATE, and θ for females (six more). The estimators for expectations are sample analogs.

Hence, all estimators can be expressed as linear combinations of sample means. For completeness, I present the fifteen descriptive estimands for a single triplet (d, d', a) below.

$$\text{Population Mean:} \quad \mu_{g,d,a} = \mathbb{E}[Y_{i,a} \mid G_i = g,\, D_i = d],$$

$$\text{DID:} \quad \begin{cases} \delta_{\text{APO}}(g,d,d',a) = \mu_{g,d,d-1} + \mu_{g,d',a} - \mu_{g,d',d-1}, \\ \delta_{\text{ATE}}(g,d,d',a) = \mu_{g,d,a} - \delta_{\text{APO}}(g,d,d',a), \\ \theta(g,d,d',a) = \dfrac{\delta_{\text{ATE}}(g,d,d',a)}{\delta_{\text{APO}}(g,d,d',a)}. \end{cases}$$

$$\text{TD:} \quad \Delta\delta_{\text{ATE}}(d,d',a) = \delta_{\text{ATE}}(f,d,d',a) - \delta_{\text{ATE}}(m,d,d',a),$$

$$\text{NTD:} \quad \Delta\delta_{\theta}(d,d',a) = \theta(f,d,d',a) - \theta(m,d,d',a),$$

$$\text{NTD Alternative Estimand:} \quad \Delta\delta_{\text{NTD-ALT}}(d,d',a) = \frac{\mu_{f,d,a}}{\mu_{m,d,a}} - \frac{\delta_{\text{APO}}(f,d,d',a)}{\delta_{\text{APO}}(m,d,d',a)},$$

$$\text{TD Under Null for Fathers:} \quad \begin{cases} \delta^{\text{TD-NULL}}_{\text{APO}}(d,d',a) = \delta_{\text{APO}}(f,d,d',a) + \delta_{\text{ATE}}(m,d,d',a), \\ \delta^{\text{TD-NULL}}_{\text{ATE}}(d,d',a) = \mu_{f,d,d-1} - \delta^{\text{TD-NULL}}_{\text{APO}}(d,d',a), \\ \theta^{\text{TD-NULL}}(d,d',a) = \dfrac{\delta^{\text{TD-NULL}}_{\text{ATE}}(d,d',a)}{\delta^{\text{TD-NULL}}_{\text{APO}}(d,d',a)}. \end{cases}$$

$$\text{NTD Under Null for Fathers :} \quad \begin{cases} \delta^{\text{NTD-NULL}}_{\text{APO}}(d,d',a) = \delta_{\text{APO}}(f,d,d',a) \times \dfrac{\mu_{m,d,a}}{\delta_{\text{APO}}(m,d,d',a)}, \\ \delta^{\text{NTD-NULL}}_{\text{ATE}}(d,d',a) = \mu_{f,d,a} - \delta^{\text{NTD-NULL}}_{\text{APO}}(d,d',a), \\ \theta^{\text{NTD-NULL}}(d,d',a) = \dfrac{\delta^{\text{NTD-NULL}}_{\text{ATE}}(d,d',a)}{\delta^{\text{NTD-NULL}}_{\text{APO}}(d,d',a)}. \end{cases}$$

Constructing estimators is straightforward. Let $S_i(g,d) = 1_{\{G_i=g,\, D_i=d\}}$ and $n_{g,d} = \sum_i S_i(g,d)$. Denote by $\bar{Y}_{g,d,a} = n_{g,d}^{-1} \sum_i S_i(g,d)\, Y_{i,a}$ the sample mean. Replacing population expectations with their corresponding sample analogs in the expressions above yields the estimators used in the empirical analysis.

Next, I turn to discuss inference. All standard errors reported in the paper are obtained using influence functions (IF). Before stating the IFs explicitly, I briefly discuss how IFs are used to quantify uncertainty in the above estimators. Let $\hat{\delta}$ denote a generic estimator of δ. Under standard regularity conditions, $\sqrt{n}\,(\hat{\delta} - \delta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_i + o_p(1)$, where φ_i is the influence function. The asymptotic variance is $\text{Var}(\hat{\delta}) = \frac{1}{n^2} \sum_{i=1}^{n} \varphi_i^2$. When data are clustered—the likely case in child-penalty applications—we replace individual observations with cluster sums. Let clusters be indexed by $c \in \{1, \ldots, C\}$, and let $\mathcal{I}_c$ denote the set of

observations in cluster c. The clustered variance is $\mathrm{Var}_{\mathrm{cluster}}(\hat{\delta}) = \frac{1}{n^2}\sum_{c=1}^{C}\left(\sum_{i\in\mathcal{I}_c}\hat{\varphi}_i\right)^2$. All standard errors in the paper implement this cluster-robust IF-based formula. Finally, the IF of a composite estimator can be obtained using the chain rule for partial derivatives, which conveniently allows the construction of IFs for the above composite estimators. Below are the IFs for the estimators defined above, for a given triplet (g, d, d', a).

For a generic estimand A, let ψ_A denote the IF of its estimator $\widehat{A}$. Using the moment–ratio identity for conditional means, $\mu_{g,d,a} = \frac{\mathbb{E}[S(g,d)\,Y_a]}{\mathbb{E}[S(g,d)]}$. From the IF of a (unconditional) mean,

$$\psi_{\mathbb{E}[S(g,d)]} = S(g,d) - \mathbb{E}[S(g,d)], \qquad \psi_{\mathbb{E}[S(g,d)\,Y_a]} = S(g,d)Y_a - \mathbb{E}[S(g,d)Y_a].$$

For a generic ratio $C = A/B$ with $B \neq 0$, the chain rule gives $\psi_C = \frac{1}{B}\,\psi_A - \frac{A}{B^2}\,\psi_B$. This yields

$$\psi_{\mu_{g,d,a}} = \frac{S(g,d)}{\mathbb{E}[S(g,d)]}\Big(Y_a - \mu_{g,d,a}\Big).$$

Using linearity and the chain rule, we obtain the following IFs.

$$
\begin{aligned}
\psi_{\delta_{\text{APO}}(g,d,d',a)} &= \psi_{\mu_{g,d,d-1}} + \psi_{\mu_{g,d',a}} - \psi_{\mu_{g,d',d-1}},\\
\psi_{\delta_{\text{ATE}}(g,d,d',a)} &= \psi_{\mu_{g,d,d-1}} - \psi_{\delta_{\text{APO}}(g,d,d',a)},\\
\psi_{\theta(g,d,d',a)} &= \frac{1}{\delta_{\text{APO}}(g,d,d',a)}\,\psi_{\delta_{\text{ATE}}(g,d,d',a)} - \frac{\delta_{\text{ATE}}(g,d,d',a)}{\delta_{\text{APO}}(g,d,d',a)^2}\,\psi_{\delta_{\text{APO}}(g,d,d',a)},\\
\psi_{\Delta\delta_{\text{ATE}}(d,d',a)} &= \psi_{\delta_{\text{ATE}}(f,d,d',a)} - \psi_{\delta_{\text{ATE}}(m,d,d',a)},\\
\psi_{\Delta\delta_{\theta}(d,d',a)} &= \psi_{\theta(f,d,d',a)} - \psi_{\theta(m,d,d',a)},\\
\psi_{\Delta\delta_{\text{NTD-ALT}}(d,d',a)} &= \frac{1}{\mu_{m,d,a}}\psi_{\mu_{f,d,a}} - \frac{\mu_{f,d,a}}{\mu_{m,d,a}^2}\psi_{\mu_{m,d,a}}\\
&\quad - \frac{1}{\delta_{\text{APO}}(m,d,d',a)}\psi_{\delta_{\text{APO}}(f,d,d',a)} + \frac{\delta_{\text{APO}}(f,d,d',a)}{\delta_{\text{APO}}(m,d,d',a)^2}\psi_{\delta_{\text{APO}}(m,d,d',a)},\\
\psi_{\delta_{\text{APO}}^{\text{TD-NULL}}(d,d',a)} &= \psi_{\delta_{\text{APO}}(f,d,d',a)} + \psi_{\delta_{\text{ATE}}(m,d,d',a)},\\
\psi_{\delta_{\text{ATE}}^{\text{TD-NULL}}(d,d',a)} &= \psi_{\mu_{f,d,d-1}} - \psi_{\delta_{\text{APO}}^{\text{TD-NULL}}(d,d',a)},\\
\psi_{\theta^{\text{TD-NULL}}(d,d',a)} &= \frac{1}{\delta_{\text{APO}}^{\text{TD-NULL}}(d,d',a)}\,\psi_{\delta_{\text{ATE}}^{\text{TD-NULL}}(d,d',a)} - \frac{\delta_{\text{ATE}}^{\text{TD-NULL}}(d,d',a)}{(\delta_{\text{APO}}^{\text{TD-NULL}}(d,d',a))^2}\,\psi_{\delta_{\text{APO}}^{\text{TD-NULL}}(d,d',a)},\\
\psi_{\delta_{\text{APO}}^{\text{NTD-NULL}}(d,d',a)} &= \frac{\mu_{m,d,a}}{\delta_{\text{APO}}(m,d,d',a)}\,\psi_{\delta_{\text{APO}}(f,d,d',a)} + \frac{\delta_{\text{APO}}(f,d,d',a)}{\delta_{\text{APO}}(m,d,d',a)}\,\psi_{\mu_{m,d,a}}\\
&\quad - \frac{\mu_{m,d,a}\,\delta_{\text{APO}}(f,d,d',a)}{\delta_{\text{APO}}(m,d,d',a)^2}\,\psi_{\delta_{\text{APO}}(m,d,d',a)},\\
\psi_{\delta_{\text{ATE}}^{\text{NTD-NULL}}(d,d',a)} &= \psi_{\mu_{f,d,a}} - \psi_{\delta_{\text{APO}}^{\text{NTD-NULL}}(d,d',a)},\\
\psi_{\theta^{\text{NTD-NULL}}(d,d',a)} &= \frac{1}{\delta_{\text{APO}}^{\text{NTD-NULL}}(d,d',a)}\,\psi_{\delta_{\text{ATE}}^{\text{NTD-NULL}}(d,d',a)} - \frac{\delta_{\text{ATE}}^{\text{NTD-NULL}}(d,d',a)}{(\delta_{\text{APO}}^{\text{NTD-NULL}}(d,d',a))^2}\,\psi_{\delta_{\text{APO}}^{\text{NTD-NULL}}(d,d',a)}.
\end{aligned}
$$

Similar to above, to construct estimators for IFs ($\hat{\psi}$) replace expectations with sample analogs. Cluster–robust standard errors are then computed using

$$
\widehat{\text{Var}}_{\text{cluster}}(\hat{A}) = \frac{1}{n^2}\sum_{c=1}^{C}\left(\sum_{i\in\mathcal{I}_c}\hat{\psi}_{A,i}\right)^2.
$$

E Appendix Covariates

E.1 Appendix DID with covariates

D.1.1. Identification. Additional assumptions that are necessary for the following identification result are:

Assumption 5 (Conditional Parallel Trends)**.** For gender g, treatment group d, control group d', target age a, and all $x \in \text{supp}(X)$, $\gamma_{\text{CPT}}(g, d, d', a, x) = 0$.

Assumption E.1 (Conditional No Anticipation)**.** For gender g, treatment group d, target age $a < d$, and all $x \in \text{supp}(X)$: $APO\left(g, d, d, d-1, x\right) = APO\left(g, d, \infty, d-1, x\right)$.

Assumption E.2 (Overlap)**.** For gender g, treatment group d, control group d' and all $x \in \text{supp}(X)$: $p(g, d, d', x) \in (0, 1)$.

Next, we introduce additional needed notation. Let

$$
\begin{aligned}
w_i^1(g, d) &= \frac{1_{\{G_i=g, D_i=d\}}}{\mathbb{E}[1_{\{G=g, D=d\}}]} \\
w_i^2(g, d, d', x) &= \frac{p(g, d, d', x)}{1 - p(g, d, d', x)} \frac{1_{\{G_i=g, D_i=d'\}}}{\mathbb{E}\left[1_{\{G_i=g, D_i=d\}}\right]} \\
\mu(g, d, d', a, x) &= \mathbb{E}[Y_a - Y_{d-1} \mid G = g, D = d', X = x]
\end{aligned}
$$

where $1_{\{A\}}$ is an indicator function equal to 1 if condition A holds and 0 otherwise, and

$$
p\left(g, d, d', x\right) = P\left(D = d \mid G = g, D \in \{d, d'\}, X = x\right)
$$

denotes the probability to be from treatment group d conditional on X. Furthermore, define the following descriptive estimands:

$$
\begin{aligned}
\delta_{\text{APO}}^{\text{OR}}(g, d, d', a) &= \mathbb{E}\left[w^1(g, d)\left(Y_{d-1} + \mu(g, d, d', a, X)\right)\right], \\
\delta_{\text{APO}}^{\text{IPW}}(g, d, d', a) &= \mathbb{E}\left[w^1(g, d) Y_{d-1} + w^2(g, d, d', X)(Y_a - Y_{d-1})\right], \\
\delta_{\text{APO}}^{\text{DR}}(g, d, d', a) &= \delta_{\text{APO}}^{\text{IPW}}(g, d, d', a) + \mathbb{E}\left[\left(w^1(g, d) - w^2(g, d, d', X)\right) \mu(g, d, d', a, X)\right].
\end{aligned}
$$

The following result shows identification is achieved for all three descriptive estimands.

Proposition E.1. *If Assumptions 5, E.1, and E.2 hold for gender g, treatment group d, target age a, and control group $d' = a + 1$, then*

$$
APO(g, d, \infty, a) = \delta_{\text{APO}}^{\text{OR}}(g, d, d', a) = \delta_{\text{APO}}^{\text{IPW}}(g, d, d', a) = \delta_{\text{APO}}^{\text{DR}}(g, d, d', a).
$$

Proof. I show the result for $\delta^{\text{DR}}_{\text{APO}}(\cdot)$; the proofs for $\delta^{\text{OR}}_{\text{APO}}(\cdot)$ and $\delta^{\text{IPW}}_{\text{APO}}(\cdot)$ follow similarly. From Theorem 1 in Callaway and Sant'Anna (2021) we obtain $ATE(g,d,a) = \mathbb{E}[w^1(g,d)Y_a] - \delta^{\text{DR}}_{\text{APO}}(g,d,d',a)$. Substituting the definitions implies $APO(g,d,d,a) - APO(g,d,\infty,a) = \mathbb{E}[w^1(g,d)Y_a] - \delta^{\text{DR}}_{\text{APO}}(g,d,d',a)$. Under consistency and Assumption E.1, $APO(g,d,d,a) = \mathbb{E}[w^1(g,d)Y_a]$, completing the proof. □

D.1.2. Estimation and Inference. I begin by deriving the influence functions for the APO. The DR identification result from Appendix E.1 can be expressed in terms of a score function. Let

$$\begin{aligned}
\psi_{\text{APO}}\left(O_i,\eta;g,d,d',a\right) &= w^1_i(g,d)\left(Y_{i,d-1} + \mu(g,d,d',a,X_i)\right) \\
&\quad + w^2_i(g,d,d',X_i)\left(Y_{i,a} - Y_{i,d-1} - \mu(g,d,d',a,X_i)\right), \\
\Psi_{\text{APO}}\left(O_i,\eta,\tau;g,d,d',a\right) &= \psi_{\text{APO}}\left(O_i,\eta;g,d,d',a\right) - w^1_i(g,d)APO\left(g,d,\infty,a\right),
\end{aligned}$$

where $O_i = (Y_{i,a}, Y_{i,d-1}, D_i, G_i, X_i)$ is the observable data, $\eta = \{p,\mu\}$ denotes the true values of the auxiliary functions, and $\tau = \{APO, ATE, \theta\}$ denotes the true values of the target causal parameters, and ψ_{APO} and Ψ_{APO} are a score function and influence function for the APO, respectively. The identification result can then be written as a moment condition $\mathbb{E}\left[\Psi_{\text{APO}}\left(O_i,\eta,\tau;g,d,d',a\right)\right] = 0$. Next, let

$$\begin{aligned}
\psi_{\text{ATE}}\left(O_i,\eta;g,d,d',a\right) &= w^1_i(g,d)Y_{i,a} - \psi_{\text{APO}}\left(O_i,\eta;g,d,d',a\right), \\
\Psi_{\text{ATE}}\left(O_i,\eta,\tau;g,d,d',a\right) &= \psi_{\text{ATE}}\left(O_i,\eta;g,d,d',a\right) - w^1_i(g,d)ATE\left(g,d,a\right),
\end{aligned}$$

be the score function and influence function of the ATE, respectively. The influence functions obey the chain rule. Hence

$$\begin{aligned}
\Psi_\theta(O_i,\eta,\tau;g,d,d',a) &= -\frac{ATE(g,d,a)}{APO(g,d,\infty,a)^2}\Psi_{\text{APO}}\left(O_i,\eta,\tau;g,d,d',a\right) \\
&\quad + \frac{1}{APO(g,d,\infty,a)}\Psi_{\text{ATE}}\left(O_i,\eta,\tau;g,d,d',a\right). \qquad \text{(E9)}
\end{aligned}$$

The estimator of $\theta(g,d,a)$ and its standard error can now be calculated as follows. First, estimate $\widehat{\psi}_{\text{APO}}$, $\widehat{APO}$ and $\widehat{\Psi}_{\text{APO}}$ using Algorithm 1. Given the estimates for the APO calcu-

Algorithm 1 Double-robust APO estimation

Input: 2-by-2 dataset, with a treated group ($D_i = d$), and closest not-yet-treated control group ($D_i = a+1$), for each unit observable data $O_i = (Y_{i,a}, Y_{i,d-1}, D_i, G_i, X_i)$

1: Partition units into $k = 1, ..., K$ folds of approximately equal size. Within fold k, train a model of $\widehat{p}_k(g, d, d', x)$ and $\widehat{\mu}_k(g, d, d', a, x)$ using observations not in fold k.

2: Denote by $k(i)$ the fold of unit i. Estimate

$$\widehat{w}_i^1(g,d) = \frac{1_{\{G_i=g, D_i=d\}}}{\mathbb{E}_{n,-k(i)}\left[1_{\{G=g,D=d\}}\right]},$$
$$\widehat{w}_i^2(g,d,d',X_i) = \frac{\widehat{p}_{k(i)}(g,d,d',X_i)}{1-\widehat{p}_{k(i)}(g,d,d',X_i)} \frac{1_{\{G_i=g,D_i=d'\}}}{\mathbb{E}_{n,-k(i)}\left[1_{\{G_i=g,D_i=d\}}\right]},$$

where $\mathbb{E}_{n,-k(i)}$ denotes the sample mean after omitting observations from fold $k(i)$

3: Estimate

$$\begin{aligned}
\widehat{\psi}_{\text{APO}}(O_i, \widehat{\eta}; g, d, d', a) &= \widehat{w}_i^1(g,d)\left(Y_{i,d-1} + \widehat{\mu}_{k(i)}(g,d,d',a,X_i)\right) \\
&\quad + \widehat{w}_i^2(g,d,d',X_i)\left(Y_{i,a} - Y_{i,d-1} - \widehat{\mu}_{k(i)}(g,d,d',a,X_i)\right), \\
\widehat{APO}(g,d,\infty,a) &= \mathbb{E}_n[\widehat{\psi}_{\text{APO}}(O_i, \widehat{\eta}; g, d, d', a)], \\
\widehat{\Psi}_{\text{APO}}(O_i, \widehat{\eta}, \widehat{\tau}; g, d, d', a) &= \widehat{\psi}_{\text{APO}}(O_i, \widehat{\eta}; g, d, d', a) - \widehat{w}_i^1(g,d)\widehat{APO}(g,d,\infty,a).
\end{aligned}$$

4: Estimate the standard errors using

$$\sqrt{\mathbb{E}_n\left[\widehat{\Psi}_{\text{APO}}(O_i, \widehat{\eta}, \widehat{\tau}; g, d, d', a)^2\right]/n}.$$

late

$$\begin{aligned}
\widehat{\psi}_{\text{ATE}}(O_i, \widehat{\eta}; g, d, d', a) &= \widehat{w}_i^1(g,d) Y_{i,a} - \widehat{\psi}_{APO}(O_i, \widehat{\eta}; g, d, d', a) \\
\widehat{ATE}(g,d,a) &= \mathbb{E}_n[\widehat{\psi}_{\text{ATE}}(O_i, \widehat{\eta}; g, d, d', a)], \\
\widehat{\Psi}_{\text{ATE}}(O_i, \widehat{\eta}, \widehat{\tau}; g, d, d', a) &= \widehat{\psi}_{\text{ATE}}(O_i, \widehat{\eta}; g, d, d', a) - \widehat{w}_i^1(g,d)\widehat{ATE}(g,d,a).
\end{aligned}$$

These estimates can be used to calculate $\widehat{\Psi}_\theta$ by substituting estimands with their respective estimators in (E9). Finally, we can calculate the estimator $\widehat{\theta}(\cdot) = \widehat{ATE}(\cdot)/\widehat{APO}(\cdot)$, and its standard error $\sqrt{\mathbb{E}_n[\widehat{\Psi}_\theta(\cdot)^2]/n}$.

E.2 Appendix TD with covariates

D.2.1. Identification. First, I introduce an additional assumption.

Assumption 6 (Equal Difference in Conditional Trends)**.** For treatment group d, control group d', target age a, and all $x \in \text{supp}(X)$, $\gamma_{\text{CPT}}(f, d, d', a, x) = \gamma_{\text{CPT}}(m, d, d', a, x)$.

Next, I introduce additional notation, building upon the notation in Appendix E.1. Let

$$\pi(g, d, d', x) = P\big((G, D) = (g, d) \mid D \in \{d, d'\}, X = x\big)$$

be the propensity to be from gender g and treatment group d among units from treatment groups $D \in \{d, d'\}$ conditional on $X = x$, and

$$w_i^3(g, g', d, d', x) = \frac{\pi(g, d, d', x)}{\pi(g', d', d, x)} \frac{1_{\{G_i = g', D_i = d'\}}}{\mathbb{E}\left[1_{\{G_i = g, D_i = d\}}\right]}$$

be a weighting function. Next, I define six descriptive estimands. The first three target the female average CATE:

$$\begin{aligned}
\delta^{\text{OR}}_{\text{ATE}-1}(d, d', a) &= \mathbb{E}_X\left[\big(w^1(f, d)\big)\left(Y_a - Y_{d-1} - \mu(f, d, d', a, X)\right)\right], \\
\delta^{\text{IPW}}_{\text{ATE}-1}(d, d', a) &= \mathbb{E}_X\left[\big(w^3(f, f, d, d, X) - w^3(f, f, d, d', X)\big)\left(Y_a - Y_{d-1}\right)\right], \\
\delta^{\text{DR}}_{\text{ATE}-1}(d, d', a) &= \mathbb{E}_X\left[\big(w^1(f, d) - w^3(f, f, d, d', X)\big)\left(Y_a - Y_{d-1} - \mu(f, d, d', a, X)\right)\right],
\end{aligned}$$

and the next three target the male average CATE, reweighted to match the female covariate distribution:

$$\begin{aligned}
\delta^{\text{OR}}_{\text{ATE}-2}(d, d', a) &= \mathbb{E}_X\left[\big(w^1(f, d)\big)\left(\mu(m, d, d, a, X) - \mu(m, d, d', a, X)\right)\right], \\
\delta^{\text{IPW}}_{\text{ATE}-2}(d, d', a) &= \mathbb{E}_X\left[\big(w^3(f, m, d, d, X) - w^3(f, m, d, d', X)\big)\left(Y_a - Y_{d-1}\right)\right], \\
\delta^{\text{DR}}_{\text{ATE}-2}(d, d', a) &= \mathbb{E}_X\left[\big(w^3(f, m, d, d, X) - w^3(f, m, d, d', X)\big)\left(Y_a - Y_{d-1}\right)\right] \\
&\quad + \mathbb{E}_X\left[\big(w^1(f, d) - w^3(f, m, d, d, X)\big)\,\mu(m, d, d, a, X)\right] \\
&\quad - \mathbb{E}_X\left[\big(w^1(f, d) - w^3(f, m, d, d', X)\big)\,\mu(m, d, d', a, X)\right].
\end{aligned}$$

Before proceeding to the statement and proof of the proposition, I prove two useful lemmas.

Lemma E.1. *If Assumptions 6, E.1, and E.2 hold for both genders $g \in \{f, m\}$, treatment group d, control group d', and target age a, then*

$$\mathbb{E}\left[\delta_{\text{CATE}}\left(m, d, d', a\right) \mid G = f, D = d\right] = \delta^{\text{OR}}_{\text{ATE}-2}(d, d', a) = \delta^{\text{IPW}}_{\text{ATE}-2}(d, d', a) = \delta^{\text{DR}}_{\text{ATE}-2}(d, d', a).$$

Proof of Lemma E.1. The proof follows directly from Lemma 1 in Leventer (2025). □

The second lemma is as follows:

Lemma E.2. *If Assumptions 6, E.1, and E.2 hold for both genders $g \in \{f, m\}$, treatment group d, control group d', and target age a, then*

$$\mathbb{E}\left[\delta_{\text{CATE}}\left(f, d, d', a\right) \mid G = f, D = d\right] = \delta^{\text{OR}}_{\text{ATE}-1}(d, d', a) = \delta^{\text{IPW}}_{\text{ATE}-1}(d, d', a) = \delta^{\text{DR}}_{\text{ATE}-1}(d, d', a).$$

Proof of Lemma E.2. The proof of the lemma follows from Theorem 1 in Callaway and Sant'Anna (2021). □

Finally, we are ready to state the TD with covariates identification result. Let

$$\vartheta(d, a) = \mathbb{E}\left[CATE\left(f, d, a, X\right) - CATE\left(m, d, a, X\right) \mid G = f, D = d\right].$$

Proposition E.2. *If Assumptions 6, E.1, and E.2 hold for both genders $g \in \{f, m\}$, treatment group d, control group d', and target age a, then*

$$\begin{aligned}
\tau(d, a) &= \delta^{\text{OR}}_{\text{ATE}-1}(d, d', a) - \delta^{\text{OR}}_{\text{ATE}-2}(d, d', a) \\
&= \delta^{\text{IPW}}_{\text{ATE}-1}(d, d', a) - \delta^{\text{IPW}}_{\text{ATE}-2}(d, d', a) \\
&= \delta^{\text{DR}}_{\text{ATE}-1}(d, d', a) - \delta^{\text{DR}}_{\text{ATE}-2}(d, d', a).
\end{aligned}$$

Proof of Proposition E.2. Combining Lemmas E.1 and E.2 completes the proof. □

D.2.2. Estimation and Inference. This appendix provides a concise articulation of the TD with covariates estimation process. For further details see Appendix E.1, which introduces the notation and discusses similar arguments for DID with covariates. Let

$$\psi_{\text{TD}}\left(O_i, \eta; g, g', d, d', a\right) = \left(w^1_i(g, d) - w^3_i(g, g', d, d', X_i)\right)\left(Y_{i,a} - Y_{i,d-1} - \mu(g', d, d', a, X_i)\right),$$

where $\eta = \{\pi, \mu\}$ denotes the true values of the auxiliary functions. Define the influence function

$$\begin{aligned}
\Psi_{\text{TD}}\left(O_i, \eta, \tau; d, d', a\right) = &\ \psi_{\text{TD}}\left(O_i, \eta; f, f, d, d', a\right) \\
&- \psi_{\text{TD}}\left(O_i, \eta; f, m, d, d, a\right) \\
&+ \psi_{\text{TD}}\left(O_i, \eta; f, m, d, d', a\right) \\
&- w^1_i\left(f, d\right)\vartheta\left(d, a\right),
\end{aligned}$$

where $\tau = \{\vartheta\}$. Estimate ψ_{TD}, ϑ and Ψ_{TD} using an algorithm similar to Algorithm 1, and calculate standard errors based on the estimated Ψ_{TD}.

F Appendix Logarithmic Functional Form

This appendix examines the validity of the DID and TD identification assumptions using logarithmic earnings. I begin with DID, and then turn to TD.

F.1 DID

Recall $\gamma_{\text{PT}}(g, d, d', a)$ is the difference in counterfactual trends between treatment group d and control group d' from gender g, between ages a and $d-1$. To differentiate between levels and logs I denote this quantity in this appendix by $\gamma_{\text{PT-LEVELS}}(g, d, d', a)$. Let $y_{i,a}(d) = \log(Y_{i,a}(d))$ denote log potential earnings, and define the difference in counterfactual log potential earnings as

$$\gamma_{\text{PT-LOGS}}(g, d, d', a) = \mathbb{E}[y_a(\infty) - y_{d-1}(\infty) \mid G = g, D = d] - \mathbb{E}[y_a(\infty) - y_{d-1}(\infty) \mid G = g, D = d'].$$

Logarithm arithmetic allows us to write this as:

$$\gamma_{\text{PT-LOGS}}(g, d, d', a) = \mathbb{E}\left[\log\left(\frac{Y_a(\infty)}{Y_{d-1}(\infty)}\right) \mid G = g, D = d\right] - \mathbb{E}\left[\log\left(\frac{Y_a(\infty)}{Y_{d-1}(\infty)}\right) \mid G = g, D = d'\right].$$

That is, parallel trends under a logarithmic function form holds if the proportional growth of earnings is the same across treatment and control.

Section 3.1 argues that parallel trends in levels likely fails due to selection on treatment timing. Is the parallel trends assumption likely to hold, given selection on treatment, in logs? To answer this, return to the selection argument discussed in the main text: later-treated individuals, with higher innate human capital, enter the labor market later due to human capital accumulation but experience steeper absolute earnings growth due to their higher human capital. Crucially, these individuals likely have *lower* absolute earnings at early ages compared to the treated group, as the earlier treated are likely more established at those ages in the labor market. The faster growth and lower initial earnings imply the parallel trends in logs is violated.

Numerical Example. Suppose we compare two individuals at age 24 (one year before treatment for $d = 25$):

- Early-treated ($d = 25$): $Y_{24} = 100$, grows to $Y_{29} = 120$
- Closest not-yet-treated control group ($d = 30$): $Y_{24} = 50$, grows to $Y_{29} = 130$

In levels, the difference for the early-treated is 20 while for the later treated is 80, a violation of parallel trends assumption in levels. For the logs specification, the differences are as follows.

- Early-treated ($d = 25$): $\log(Y_{29}/Y_{24}) = \log(1.20) \approx 0.18$.
- Closest not-yet-treated control group ($d = 30$): $\log(Y_{29}/Y_{24}) = \log(2.60) \approx 0.96$.

The above shows that the proportional growth rate for later-treated individuals is larger as well. This occurs precisely because they start from a lower baseline while experiencing higher absolute growth. Hence, both the steeper earning trajectories in absolute terms and the initial lower levels of absolute earnings contribute to the violation of parallel trends in logs.

F.2 TD

Section 3.2 connects the triple differences assumption in levels to counterfactual gender inequality. I now turn to how this is related to the logarithmic specification. Recall that TD in levels assumes $\gamma_{\text{PT-LEVELS}}(f, d, d', a) = \gamma_{\text{PT-LEVELS}}(m, d, d', a)$. Similarly, TD in logs assumes $\gamma_{\text{PT-LOGS}}(f, d, d', a) = \gamma_{\text{PT-LOGS}}(m, d, d', a)$. Next, recall that $\rho(d, \infty, a)$ is the gender ratio of mean counterfactual earnings for treatment group d at age a. For this appendix, we denote this quantity as $\rho_{\text{LEVELS}}(d, \infty, a)$. Similarly, let

$$\rho_{\text{LOGS}}(d, \infty, a) = \mathbb{E}[y_a(\infty) \mid G = f, D = d] - \mathbb{E}[y_a(\infty) \mid G = m, D = d]$$

be the difference in average log counterfactual earnings between females and males. The TD assumption under the log functional form can be re-written as:

$$\begin{aligned}&\gamma_{\text{PT-LOGS}}(f, d, d', a) - \gamma_{\text{PT-LOGS}}(m, d, d', a) \\ &= \rho_{\text{LOGS}}(d, \infty, a) - \rho_{\text{LOGS}}(d, \infty, d-1) - \left[\rho_{\text{LOGS}}(d', \infty, a) - \rho_{\text{LOGS}}(d', \infty, d-1)\right].\end{aligned}$$

This shows that TD assumption in logs is equivalent to stating that the trend in log gender inequality from age $d-1$ to age a is the same across treatment groups d and d'.

Furthermore, the TD assumption under the levels and logs functional forms may be connected by assuming a specific earnings distribution. Under a log-normal earnings distribution, this statement is related to the life-cycle variation in $\rho_{\text{LEVELS}}(d, \infty, a)$. Suppose earnings are log-normally distributed with equal variance across genders. That is, assume

$$Y_a(\infty) \mid G = g, D = d \sim \text{LogNormal}\left(\mu_{g,d,a}, \sigma^2_{d,a}\right).$$

Under this assumption, we obtain $\rho_{\text{LOGS}}(d, \infty, a) = \log(\rho_{\text{LEVELS}}(d, \infty, a))$, since

$$\rho_{\text{LVLS}}(d, \infty, a) = \exp\left(\mu_{f,d,a} - \mu_{m,d,a} + \frac{\sigma^2_{d,a} - \sigma^2_{d,a}}{2}\right) = \exp\left(\mu_{f,d,a} - \mu_{m,d,a}\right).$$

Figure 2 in Section 3.2 documents life-cycle patterns of $\rho_{\text{LEVELS}}(d, \infty, a)$ across treatment groups, highlighting differences in levels but overall similar trends. Combining this descriptive evidence with the above arguments suggests that TD in logs is potentially suited for identification.

To validate whether this holds in the data, I implement pre-treatment validation tests empirically, similar to the discussion in Section 3.3. Using the estimates for counterfactual gender inequality in pre-treatment periods, presented in Section 3.2, I calculate $\log(\widehat{\rho}_{\text{LEVELS}}(d, \infty, a))$. Then, for treatment group d I calculate the trend in gender inequality as:

$$\Delta \log(\widehat{\rho}_{\text{LEVELS}}(d, a)) = \log(\widehat{\rho}_{\text{LEVELS}}(d, \infty, a)) - \log(\widehat{\rho}_{\text{LEVELS}}(d, \infty, d-1)).$$

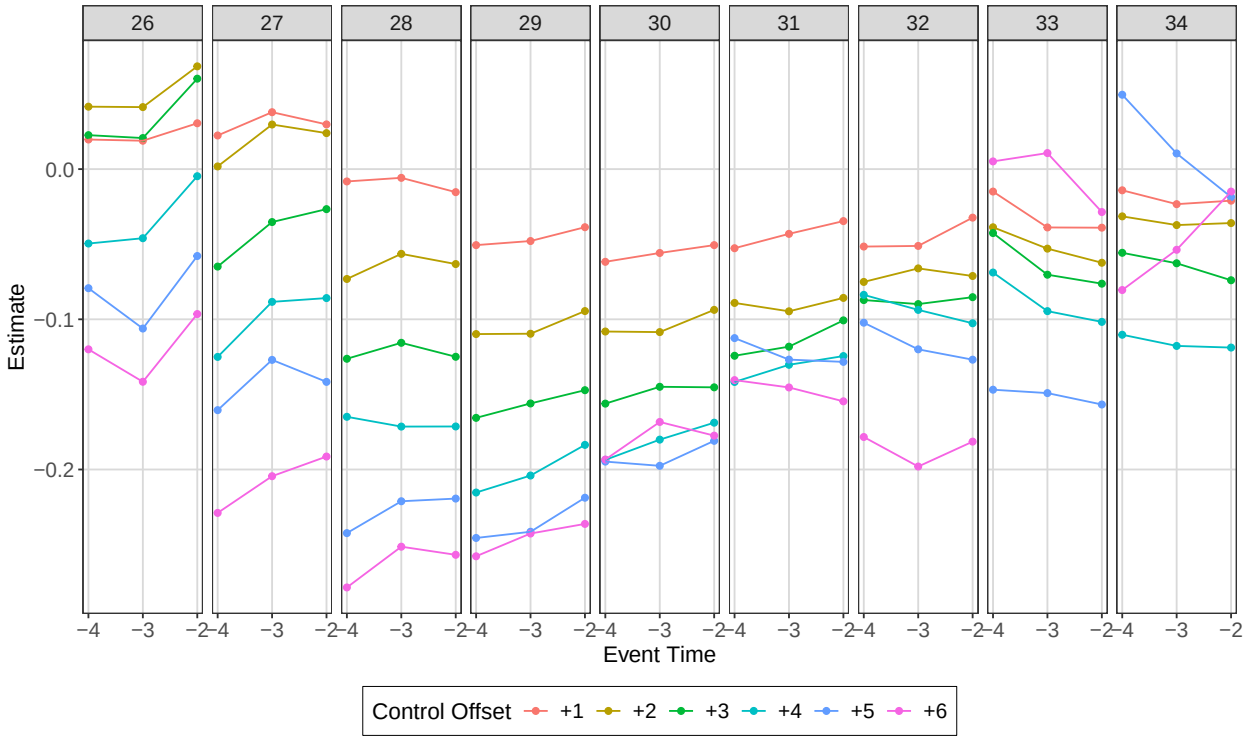
Finally, I calculate the difference in gender inequality trends between treatment d and control d' as:

$$\Delta\Delta \log(\widehat{\rho}_{\text{LEVELS}}(d, d', a)) = \Delta \log(\widehat{\rho}_{\text{LEVELS}}(d, a)) - \Delta \log(\widehat{\rho}_{\text{LEVELS}}(d', a)).$$

If TD in logs holds, then $\Delta\Delta \log(\widehat{\rho}_{\text{LEVELS}}(d, d', a))$ should be close to zero for all pre-treatment ages $a < d$. Appendix Figure F1 presents the results, documenting $\Delta\Delta \log(\widehat{\rho}_{\text{LEVELS}}(d, d', a))$ by treatment groups in columns and control groups in colors. Control groups are presented using the age-gap from treatment similar to the validation tests presented in Figure 3. E.g., $+1$ for treatment group $d = 27$ is control group $d' = 27 + 1 = 28$. The figure shows that for most treatment groups the magnitude of violations increases monotonically with the age gap between treatment and control. Also of interest, it seems that within each treatment-control pair, violations at event times -4, -3, and -2 are similar in magnitude, i.e., there is little trend. Hence, these results can be seen as suggestive that the TD assumption in log is not valid.

The results suggest that selection on treatment timing is correlated with how gender inequality evolves over the life-cycle. Early- and late-childbearing parents not only have different earnings trajectories, which introduces problems for parallel trends, but these trajectories also exhibit different gender gap dynamics.

Figure F1: Validation Tests TD in Logs



Notes: The figure presents a pre-treatment validation test for triple differences in logarithmic functional form. Treatment groups in columns and control groups in colors. For further details on treatment-control pairs see notes of Figure 3. The estimates are for $\Delta\Delta \log(\widehat{\rho}_{\text{LEVELS}}(d, d', a))$.

F.3 Conclusion

This appendix demonstrates that using logarithmic earnings instead of levels does not resolve the identification challenges discussed in Section 3. The parallel trends assumption fails in logs for the same selection reasons as in levels, with the violation potentially even larger in proportional terms. While Figure 2 combined with a log-normal assumption on the distribution on earnings suggests that TD in logs might be valid in the child penalty context, pre-treatment validation tests reveal systematic violations that worsen as the treatment-control gap increases.

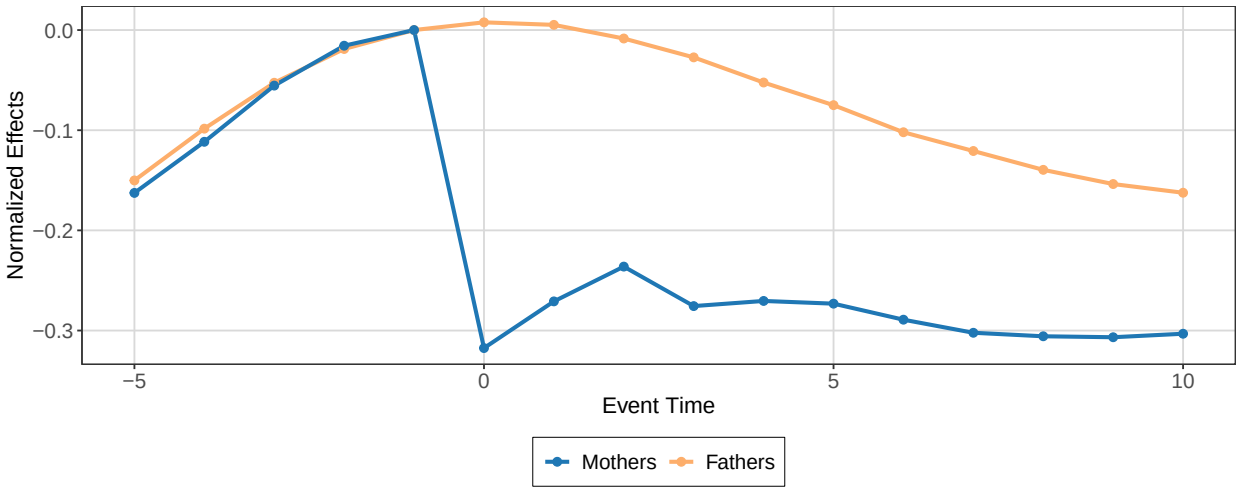
G Appendix Tables and Figures

Table G1: TD Decomposition and APO Threshold Values

		Term 1 (d, a)		Term 2 $(d, d-1)$			Term 3 (d', a)			Term 4 $(d', d-1)$		
d	a	$\widehat{\rho}$	$APO^{\star}$	$\widehat{\rho}$	$\widehat{APO}$	$=$	$\widehat{\rho}$	$\widehat{APO}$	$=$	$\widehat{\rho}$	$\widehat{APO}$	$=$
24	27	0.98	168392	0.93	40339	2690	0.90	77074	7840	0.84	42101	-6594
24	28	0.94	103394	0.93	40339	2690	0.90	88004	9031	0.86	39660	-5657
24	29	0.88	120227	0.93	40339	2690	0.85	104116	15983	0.87	37950	-4787
25	28	0.95	100921	0.90	50353	4846	0.90	88004	9031	0.82	46963	-8499
25	29	0.89	119931	0.90	50353	4846	0.85	104116	15983	0.83	44637	-7552
25	30	0.84	131542	0.90	50353	4846	0.80	117669	23089	0.84	42658	-6777
26	29	0.87	105560	0.91	58339	5058	0.85	104116	15983	0.86	51707	-7346
26	30	0.82	122592	0.91	58339	5058	0.80	117669	23089	0.88	48726	-6059
26	31	0.78	135170	0.91	58339	5058	0.78	128207	28784	0.90	46335	-4689

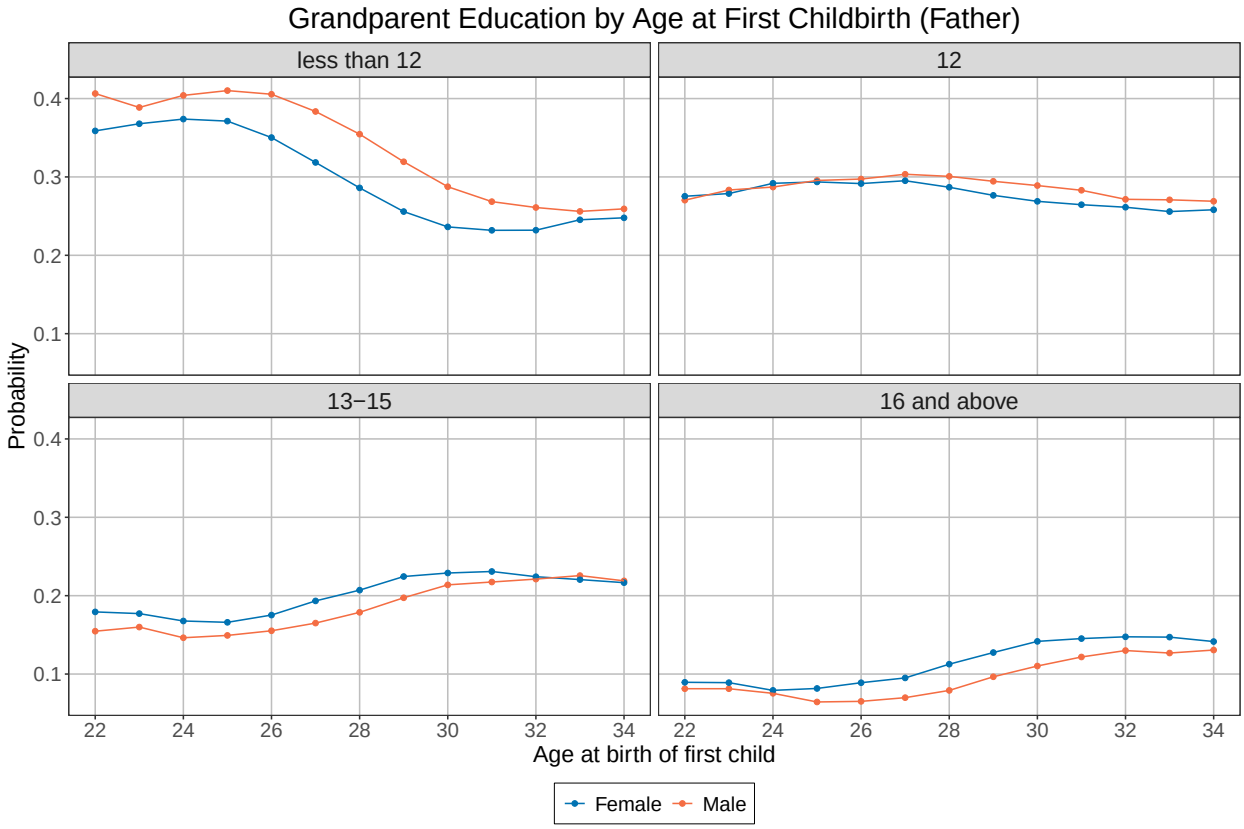
Notes: The table reports the four terms from the TD decomposition in (6). Each row corresponds to a treatment group d and a post-birth age a, with control group $d' = a + 1$ and baseline age $d - 1$. The ρ and APO components in Terms 2–4 are estimated using observed mean earnings. Columns labeled $=$ display the product $(\rho - 1)APO$; for Terms 2 and 3, the sign is reversed in accordance with (6). Term 1 reports the imputed counterfactual gender ratio $\widehat{\rho}(d, a)$, projected from the demeaned trends observed in treatment groups 29-35. $APO^{\star}$ denotes the threshold value of male counterfactual earnings at which the sum of the four terms equals zero.

Figure G1: Event Study based Child Penalty Estimates



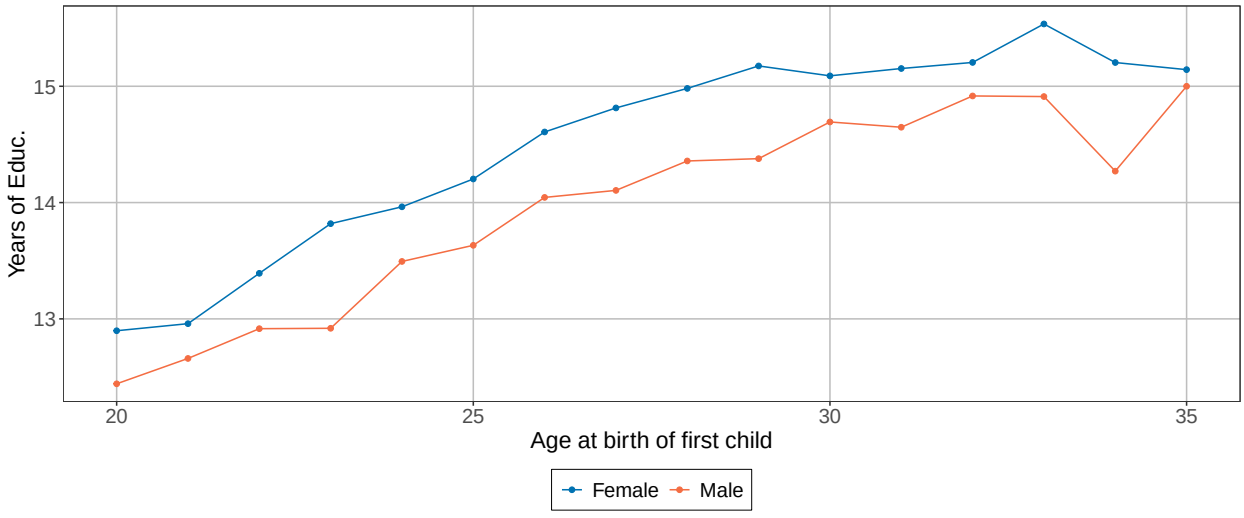
Notes: The figure presents event study based child penalty estimates. The sample is based on treatment groups $D \in \{20, \ldots, 40\}$, and only includes individuals observed in all ages from 5 years prior treatment to 10 years post treatment. For the empirical strategy see Section 1.1.

Figure G2: Grandfathers education and age at first childbirth



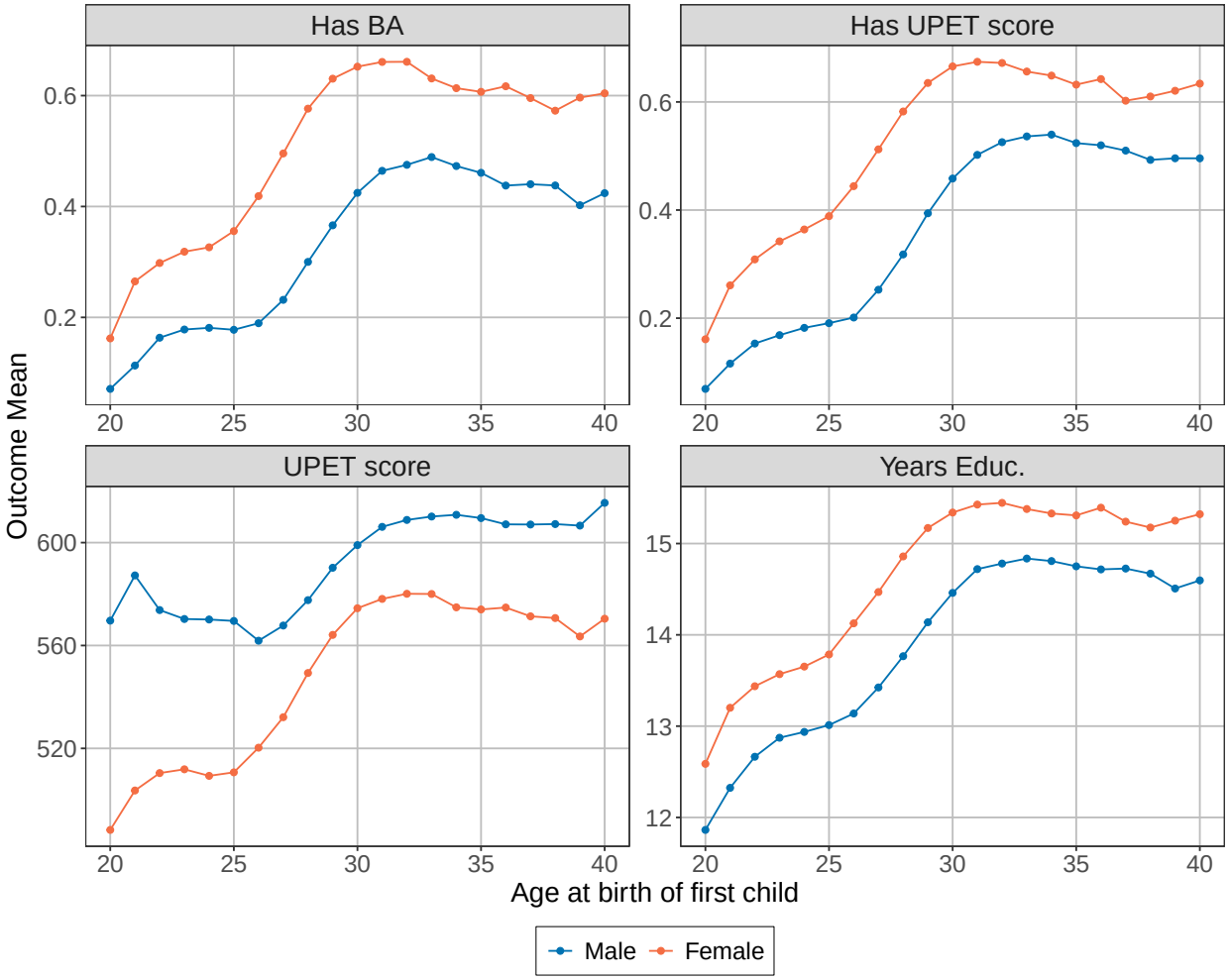
Notes: The figure presents shares of education attainment of grandfathers by age at first childbirth of the parents.

Figure G3: Years of education and age at first childbirth in PSID



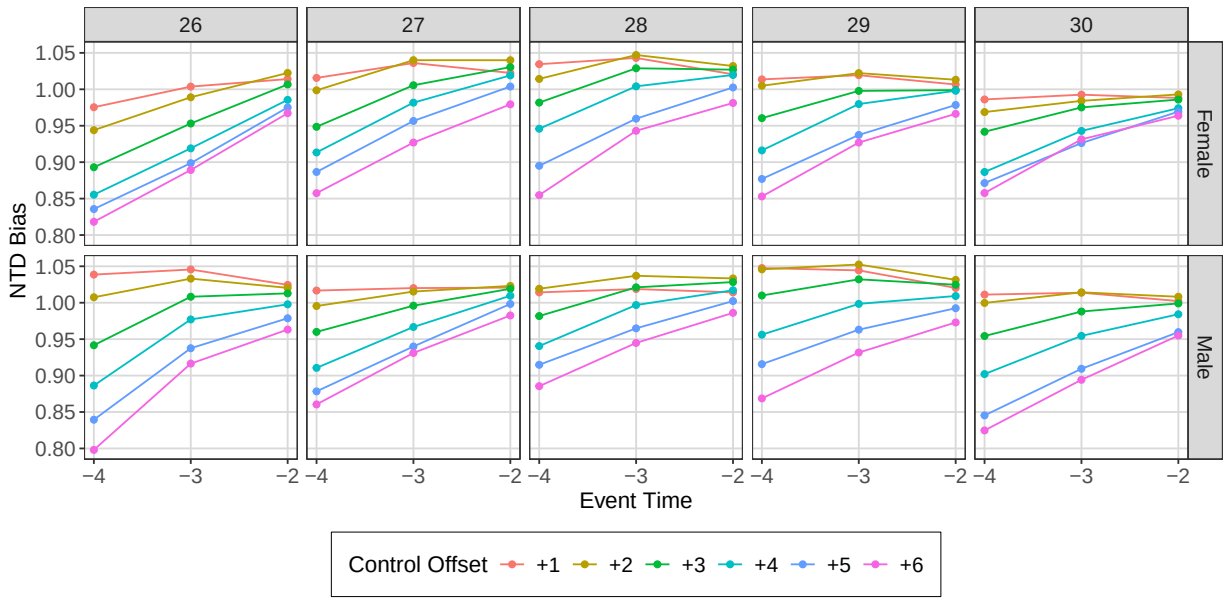
Notes: The figure presents mean years of education by age of first childbirth in the USA using the PSID dataset. The PSID dataset was taken from Cortés and Pan (2023).

Figure G4: Education and age at first childbirth



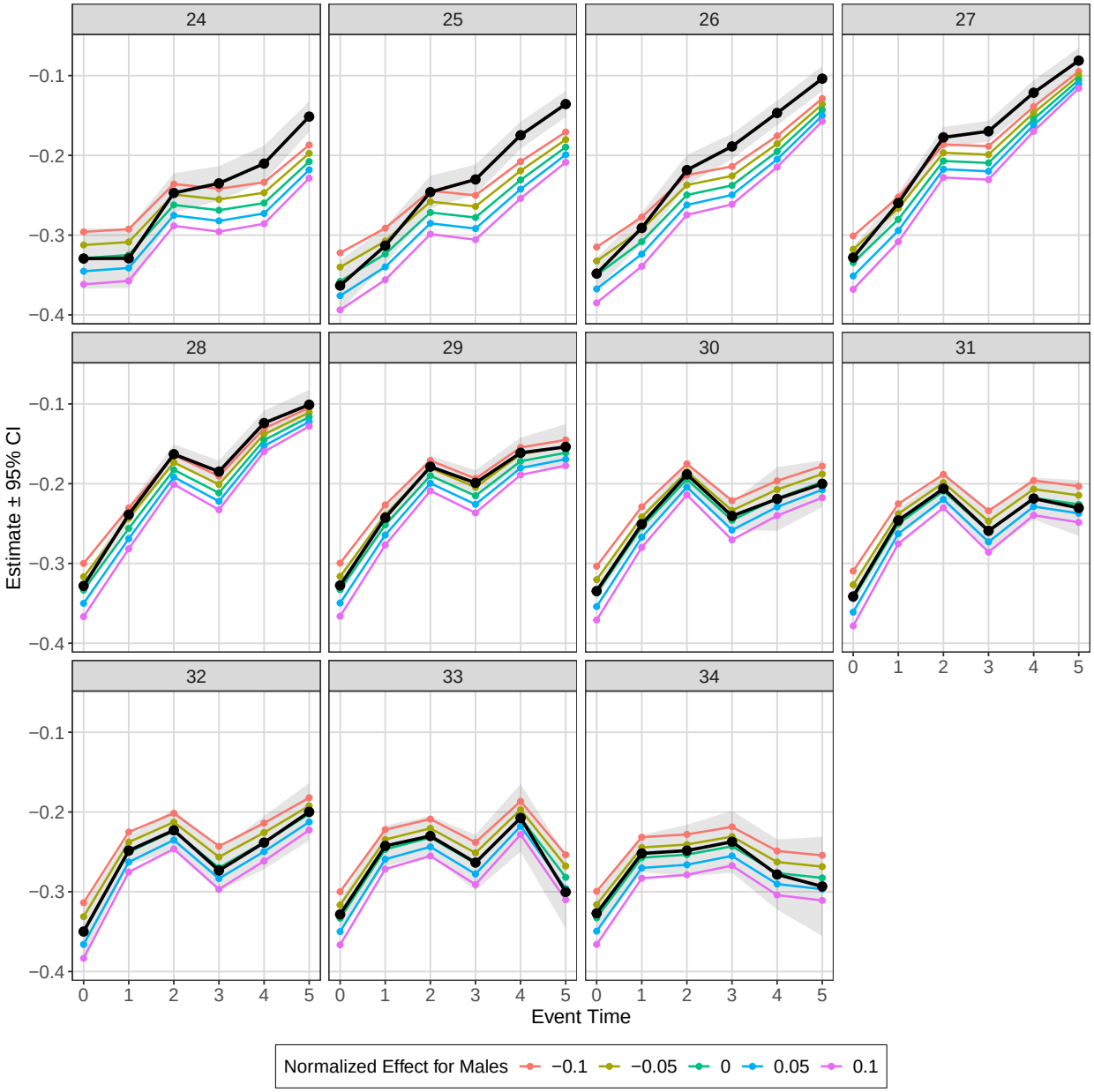
Notes: The figure presents means of education variables by the age at which an individual became a parent. The variables are reported in the title of each panel. 'Has BA' represents a dummy variable that indicates whether an individual has a college or university degree. 'Has UPET score' represents a dummy variable the indicates whether an individual has a non-missing UPET score. 'UPET score' represents the UPET score from the highest scoring test of an individual with a non-missing score. 'Years Educ.' represents the years of education.

Figure G5: Estimated Bias in Gender Gaps of Normalized Effects Under NTD



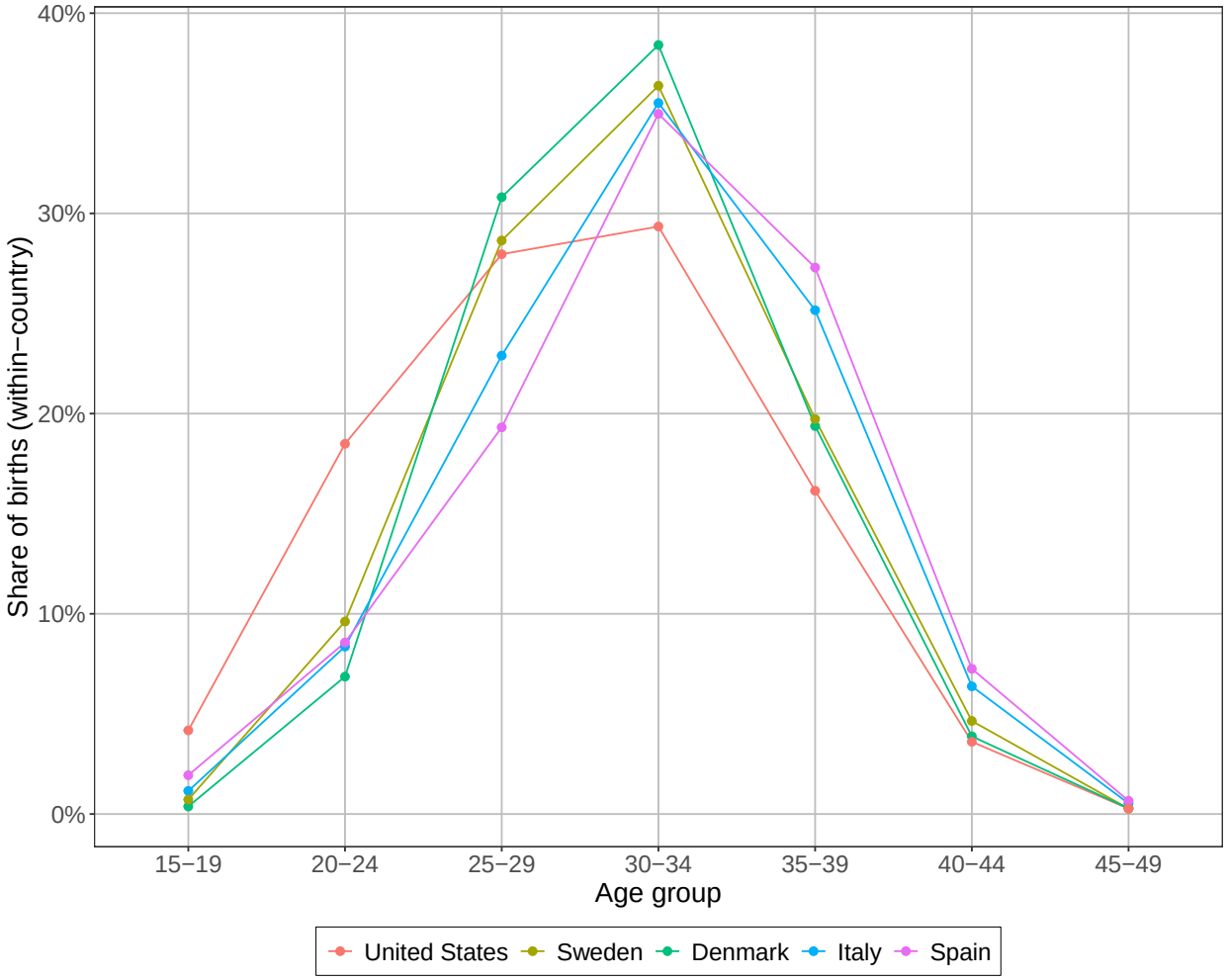
Notes: The figure reports pre-treatment estimates of the bias in gender gaps of normalized effects under the NTD framework as discussed in Theorem 1 (termed $\text{Bias}(d, d', a)$). Columns correspond to treatment groups and rows to gender. Under no anticipation (Assumption NA) in pre-treatment periods the counterfactual APO is identified via observed means, and the violation of parallel trends is identified via DID equivalent of the ATE. Hence, the bias can be estimated in pre-treatment periods using $\widehat{\text{Bias}}(d, d', a) = \frac{\mathbb{E}_n[Y_a|G=g, D=d]}{\mathbb{E}_n[Y_a|G=g, D=d] - \widehat{\delta}_{\text{ATE}}(g, d, d', a)}$ for gender $g \in \{f, m\}$ where, under NTD, the bias should be equal across genders. Colors represent different control groups used in post-treatment periods, see the notes to Figure 3 for further details.

Figure G6: Robustness to Alternative Assumptions on Male Effects



Notes: The figure presents bias-corrected NTD estimates under alternative assumptions about the normalized treatment effect for males (facets) by treatment group (columns). The black line shows baseline NTD estimates with 95% confidence intervals. Colored lines show bias-corrected estimates assuming specific values of θ_m ranging from -0.1 to 0.1, building on the discussion in Section 4.2.

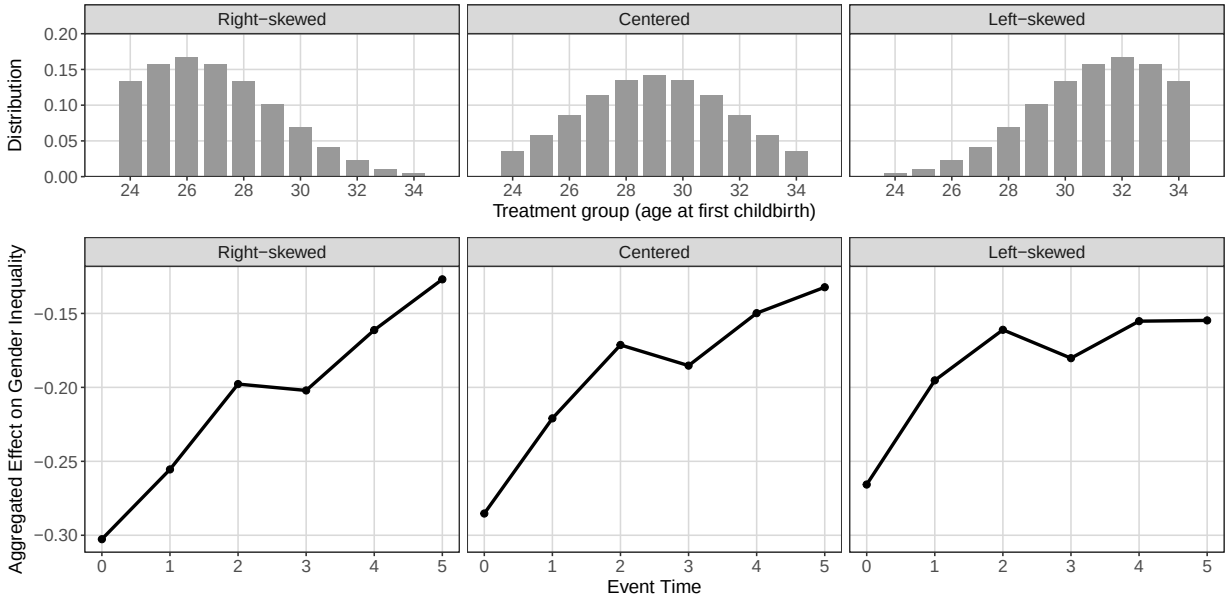
Figure G7: OECD Age at First Childbirth Distributions



Source: OECD Family Database (Indicator SF2.3).

Notes: The figure shows the distribution of age at first childbirth for selected OECD countries in 2021. The OECD database reports fertility rates for first births by five-year age groups (15–19, 20–24, ..., 45–49). For each country, the age-group fertility rates are normalized to sum to one, so the vertical axis represents the within-country share of first births across age groups.

Figure G8: Aggregated Effects under Three Hypothetical Treatment Distributions



Notes: The figure shows estimates of the aggregated effect of parenthood on the gender earnings ratio across treatment groups $D \in [24, 34]$ by treatment distribution (columns). Single-treatment-group estimates were calculated on the Israeli administrative data, as in Section 4.1. The three treatment distributions are based on normal distributions with standard deviation 3 and means 27, 29 and 31, titled Right-skewed, Centered and Left-skewed, respectively. Aggregated estimates report weighted averages of single-treatment group estimates using the density of each treatment group as weights.